Cancer risk assessment using quantitative imaging features from solid tumors and surrounding structures

Uthoff, Johanna Mariah

Uthoff, J. M. (2021). Cancer risk assessment using quantitative imaging features from solid tumors and surrounding structures [University of Iowa]. https://doi.org/10.17077/etd.66kk-u2b1

CANCER RISK ASSESSMENT USING QUANTITATIVE IMAGING FEATURES FROM SOLID TUMORS AND SURROUNDING STRUCTURES

by

Johanna Mariah Uthoff

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Biomedical Engineering in the
Graduate College of
The University of Iowa

May 2019

Thesis Supervisor: Assistant Professor Jessica C. Sieren

To Evie.

ACKNOWLEDGEMENTS

The work described here was funded in part by the American Lung Association's Lung Health Dissertation Award (LH-574107), the University of Iowa Graduate College's Post Comprehensive Research Fellowship, and the Synodos for NF1 program at the Children's Tumor Foundation.

My journey through graduate study to this thesis would not have been possible without the support and guidance of many.

To my thesis advisor, for providing unwavering encouragement, for pushing me to do more, and for providing for me the resources to prosper: Dr. Jessica Sieren.

To my thesis committee members whose guidance and careful consideration have been instrumental in advancing this work: Drs. Mona Garvin, Eric Hoffman, Joseph Reinhardt, and Milan Sonka.

To the many collaborators fostering this work: Sarah Bell, , Dr. Thomas Grosse Dr. Richard Hoffman, Dr. Prashant Nagpal, Dr. John Newell Jr., Dr. Rolando Sanchez, and Dr. Ann Schwartz.

To the lab members, past and present, you have been a source of joy, solace, and encouragement: Timothy Dougherty, Alexandra Judisch Gogola, Dr. Jacob Herrmann, Dr. Krishna Iyer, Jared Larson, Kevin Knoernschild, and Kelly Stark.

Furthermore, to the undergraduate students who have provided support and helped me grow: Frank De Stefano, Nicholas Koehn, Madhuvanthi Muralidharan, and Kimberly Schroeder.

To my friend and colleague, Dr. Emily Hammond, who listened to me, supported me, allowed me to commandeer her couch, and befriended me throughout it all.

And to my predecessor, undergraduate mentor, and friend, Dr. Samantha K.N. Dilger. The work composed here and the woman presenting it would not have been possible without her foundation and my graduate career would not have been possible without her support and friendship.

ABSTRACT

Medical imaging is a powerful tool for clinical practice allowing in-vivo insight into a patient's disease state. Many modalities exist, allowing for the collection of diverse information about the underlying tissue structure and/or function. Traditionally, medical professionals use visual assessment of scans to search for disease, assess relevant disease predictors and propose clinical intervention steps. However, the imaging data contain potentially useful information beyond visual assessment by trained professional. To better use the full depth of information contained in the image sets, quantitative imaging characteristics (QICs), can be extracted using mathematical and statistical operations on regions or volumes of interests. The process of using QICs is a pipeline typically involving image acquisition, segmentation, feature extraction, set qualification and analysis of informatics. These descriptors can be integrated into classification methods focused on differentiating between disease states. Lung cancer, a leading cause of death worldwide, is a clear application for advanced in-vivo imaging based classification methods.

We hypothesize that QICs extracted from spatially-linked and size-standardized regions of surrounding lung tissue can improve risk assessment quality over features extracted from only the lung tumor, or nodule, regions. We require a robust and flexible pipeline for the extraction and selection of disease QICs in computed tomography (CT). This includes creating an optimized method for feature extraction, reduction, selection, and predictive analysis which could be applied to a multitude of disease imaging problems. This thesis expanded a developmental pipeline for machine learning using a large multicenter controlled CT dataset of lung nodules to extract CT QICs from the nodule, surrounding parenchyma, and greater lung volume and explore CT feature interconnectivity. Furthermore, it created a validated pipeline that is more computationally and time efficient and with stability of performance. The modularity of the optimized pipeline facilitates broader application of the tool for applications beyond CT identified pulmonary nodules.

We have developed a flexible and robust pipeline for the extraction and selection of Quantitative Imaging Characteristics for Risk Assessment from the Tumor and its Environment (QIC-RATE). The results presented in this thesis support our hypothesis, showing that classification of lung and breast tumors is improved through inclusion of peritumoral signal. Optimal performance in the lung application achieved with the QIC-RATE tool incorporating 75% of the nodule diameter equivalent in perinodular parenchyma with a development performance of 100% accuracy. The stability of performance was reflected in the maintained high accuracy (98%) in the independent validation dataset of 100 CT from a separate institution. In the breast QIC-RATE application, optimal performance was achieved using 25% of the tumor diameter in breast tissue with 90% accuracy in development, 82% in validation. We address

the need for more complex assessments of medically imaged tumors through the QIC-RATE pipeline; a modular, scalable, transferrable pipeline for extracting, reducing and selecting, and training a classification tool based on QICs. Altogether, this research has resulted in a risk assessment methodology that is validated, stable, high performing, adaptable, and transparent.

PUBLIC ABSTRACT

Cancer is one of the leading causes of death worldwide. Medical imaging of tumors is an important step in the detection and diagnosis of cancer. In the lung, imaging has become a powerful tool in detecting small tumors; even still, not all detected tumors are cancerous and invasive procedures to obtain diagnostic truth carry risks. Therefore, medical professionals are often faced with the challenging task of assigning risk to a subject based primarily on the appearance of the tumor on imaging along with the subject's previous history of clinical risk-factors (age, smoking, etc.). Imaging data contains potentially useful information beyond that which is visually perceived by trained medical professionals, when image data is analyzed using computer algorithms. Here we develop a pipeline method using artificial intelligence (AI) to assign cancer risk scores that can be used by radiologists to better understand cancer risk in patients with tumors. We hypothesize that the tumor and the tissue surrounding the tumor have characteristics that differ between cancer and non-cancer tumors. Informative features are automatically calculated from the image of the tumor. We developed a method to intelligently select from all available features, those most important characteristics for distinguishing cancer from non-cancer cases. These selected features are used to teach an AI program to assign a cancer risk score to the subject's tumor. We have applied this pipeline to tumors in the lung and the breast, showing high performance in the AI program that incorporates features from the surrounding tissue. Furthermore, this method requires little human interaction to be applied which is ideal for use in clinics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITIONS OF ABBREVIATIONS

ANN	Artificial neural network
AUC-PR	Area-under-the-curve precision recall
AUC-ROC	Area-under-the-curve of the receiver-operator characteristic
BI-RADS	Breast Imaging Reporting And Data System
BTS	British Thoracic Society
BU	Brock University
CBIS-DDSM	Curated Breast Imaging Subset of the Digital Database for Screening Mammography
COPD	Chronic obstructive pulmonary disease
COPDGene	Genetics and Epidemiology of Chronic Obstructive Pulmonary Disease
CT	Computed tomography
CV	Coefficient of variation
ENNs	Ensemble of artificial neural networks
FIJI	FIJI-ImageJ
FWHM	Full-width-at-half-maximum
GC	Graph-cuts
GLRL	Run Gray Level Length Texture
GLSZ	Gray Level Size Zone Texture
ICC	Interclass correlation coefficient
IH	Intensity Histogram
INHALE	Genetic Epidemiology of Lung Cancer
IO	Information Optimization
IRB	Institutional review board
ITK-S	ITK-Snap

JD	Jaccard Distance
kCV	K-fold cross validation
LASSO	Least absolute shrinkage and selection operator
LDCT	Low-dose computed tomography
LIDC	Lung Imaging Database Consortium
LSML	Level-set maximizing likelihood
LTEM	Law's texture energy measures
LUNG-RADS	American College of Radiology's Lung Imaging Reporting and Data System
MAD	Median absolute deviation
MC	Mayo Clinic
ML	Mukhopadhyay-MatLab
MPMs	Mathematical prediction models
MR	Magnetic resonance
MVL	MeVisLab
NGTD	Neighborhood gray tone difference
NLST	National Lung Screening Trial
NSCLC	Non-small cell lung cancer
OR	Odds ratio
PET	Positron emission tomography
PFT	Pulmonary function testing
PU	Peking University
qCT	Quantitative computed tomography
QIBA	Quantitative Imaging Biomarker Alliance
QIC-RATE	Quantitative Imaging Characteristics of Risk Assessment from the Tumor and Environment
QIC	Quantitative imaging characteristic

ROI	Region of interest
SHD	Standardized Hausdroff Distance
SzSp	Measures of size and volumetric shape
TCIA	The Cancer Imaging Archive
TP_F	Final imaging encounter before diagnosis
TP_I	Initial (incidental) imaging encounter
UIHC	University of Iowa Hospitals and Clinics
VA	U.S. Department of Veterans Affairs
VER	Volumetric Error Rate

CHAPTER 1: INTRODUCTION

Medical imaging is a powerful tool for clinical practice allowing in-vivo insight into a patient's disease state. Many modalities exist including computed tomography (CT), mammography, magnetic resonance (MR), and positron emission tomography (PET). Each of these systems provides a different image rendering of tissue characteristics using unique methods of data acquisition and image construction allowing the collection of diverse information about the underlying tissue structure and/or function. Traditionally, medical professionals use visual assessment of scans to search for disease, to assess relevant disease predictors, and to propose clinical intervention steps. However, these data contain potentially useful information beyond visual assessment by trained professional. To better use the full depth of information contained in the image sets, quantitative imaging characteristics (QICs), can be extracted using mathematical and statistical operations on regions or volumes of interests.

QICs have become an area of exponential growth in research during the last thirty years[1,2]. The process of using QICs is a pipeline typically involving image acquisition, segmentation, feature extraction, set qualification and analysis of informatics. These descriptors can be integrated into classification methods focused on differentiating between disease states. This pipeline can be applied to a multitude of medical imaging problems including disease risk assessment, progression probability, and prediction of survival[3]. Cancer, a leading cause of death worldwide, is a clear application for advanced in-vivo imaging based classification methods. This dissertation focuses on the most deadly cancer – lung cancer[4].

Lung cancer nodules are histologically heterogeneous, containing a complex intermixing of cancerous cells, with regions of inflammation, fibrosis, and necrosis. Pathology-driven investigations of the surrounding perinodular parenchyma has demonstrated the difference in cellular infiltration between lung cancer and benign processes as well as the increased parenchyma distention from larger nodules[5,6]. Medical imaging presents the ability to non-invasively capture whole-tumor characteristics and quantitative CT (qCT) metrics are a way of measuring tumor heterogeneity[7-9]. Prior published research demonstrates that qCT shape and texture features extracted from CT representation of suspicious pulmonary nodules can aid in the classification of malignant versus benign and previously our group has indicated the predictive potential of including the immediately surrounding lung tissue, or parenchyma, in the analysis of solid pulmonary nodules[10-25].

We require a robust and flexible pipeline for the extraction and selection of disease characteristics in medical imaging data. This includes creating an optimized method for feature extraction, reduction, selection, and predictive analysis which could be applied to a multitude of disease imaging problems. This thesis seeks to expand a developmental pipeline for machine learning using a large multicenter

controlled CT dataset of lung nodules to extract qCT QICs from the nodule, surrounding parenchyma, and greater lung volume and explore qCT feature interconnectivity. Furthermore, it seeks to create a validated pipeline that is more computationally and time efficient and with stability of performance. We hypothesize that QICs extracted from spatially-linked and size-standardized regions of surrounding lung tissue can improve risk assessment quality over features extracted from only the lung nodule regions. Furthermore, we believe the resultant approach has predictive value (and stability) for applications beyond CT identified pulmonary nodules.

This dissertation will begin with acknowledgement of the datasets used (**Chapter 2: Datasets)** followed by an assessment of classification models previously developed using human-provided variables for lung nodule risk prediction (**Chapter 3: Mathematical Prediction Models)**. This will be followed by a study of the necessary image segmentation to systematically create VOIs (**Chapter 4: Segmentation**). We describe the development of a radiomics-based pipeline for transparent feature set identification and classifier training using Quantitative Imaging Characteristics for Risk Assessment from the Tumor and its Environment (QIC-RATE) for distinction between malignant and benign lung nodules (**Chapter 5: QIC-RATE**). We then demonstrate the applicability of the QIC-RATE pipeline to other disease distinctions (**Chapter 6: Application of QIC-RATE to Histoplasmosis Classification, Chapter 8: Application of QIC-RATE to Breast Tumor Classification**) and additional feature extraction techniques (**Chapter 7: Application of QIC-RATE to Global Lung Measures**).

CHAPTER 2: DATASETS

This dissertation focuses on x-ray-based imaging, specifically computed tomography (CT) of the lungs. CT is the standard imaging modality for pulmonary disease assessment due to its (1) fast acquisition incurring less motion artifact from cardiac and breathing rhythms, (2) volumetric high-resolution CT acquisition, and (3) high inherent contrast of lung tissue and structures. In the National Lung Screening Trial (NLST), low-dose CT (LDCT) was shown to reduce lung cancer mortality by 20% over chest radiograph[26]. Lung imaging using CT is recommended for the screening of individuals with a high-risk of lung cancer[27-33], the follow-up on tumors discovered during screening[27-33], and the typical recommended follow-up method for incidentally discovered nodules in clinic[31,34].

The applications developed in this dissertation have applicability to other tumor imaging, as a demonstration we have made appropriate modifications to apply the same technique to breast cancer mammography. Mammography has been the standard imaging modality for breast screening for due to its (1) fast acquisition, (2) low radiation exposure (~30 kVp), and (3) relatively low-cost option. The Breast Imaging Reporting And Data System (BI-RADS) was developed for a standard method of reading and reporting abnormal findings on breast mammography exams[35].

This chapter details the CT and mammography datasets used in this dissertation document. Here, we explain the original purpose(s) for collection of the origin datasets and eligibility and/or inclusion criteria. For all datasets, diagnostic truth determination method for malignancy was confirmed with histopathology via needle biopsy or surgical resection; for benign cases, truth was determined by histopathology from surgical resection or by imaging from resolution or stability for greater than two years. **Table 2.1** contains a summary of key demographic and clinical from each dataset. In each of the subsequent chapters, the numbers and relevant demographical/scanning information is included for the portion of the cohorts used.

2.1. Clinical Datasets

Establishing a retrospective clinical cohort has the advantage of access to an extensive amount of demographic, clinical, and diagnostic information. The limitation of this dataset, being retrospectively collected, is substantial heterogeneity in the CT acquisition parameters. The main source of variation in protocol is due to incidental lung nodule discovery. In these cases, the key clinical indicator requiring the order for thoracic imaging (i.e. respiratory symptoms, cardiac, trauma/emergency, referral to oncology from external physician, etc.) affects the type of CT protocol parameters applied for acquisition. Additionally, as CT scanners are updated, parameters such as reconstruction kernel type, image resolution (x-, y-, z-plane), and noise properties related to advanced hardware can impact image quality[36].

Table 2.1: Overview summary of available data that is used in this thesis.

	UIHC	Multicenter Studies			Publicly Available Dataset			
	Clinical	**COPDGene**	**INHALE**	**NLST**	**LungX**	**LIDC**	**FDA**	**DDSM**
Diagnosis	199 190M:90B	269 30M:239B	100 50M:50B	14 6M:8B	80 38M:42B	12 12M:0B	NA	1115 568M:547B
Segmentation	199	NA	NA	NA	NA	12	7	1115
Apollo	NA	213	100	14	NA	NA	NA	NA
Age (years)	60±12	62±13	63±11	65±12	61±13	64±10	NA	52±7
Sex	111X:88Y	164X:105Y	66X:34Y	7X:7Y	42X:38Y	4X:8Y	NA	1115X:0Y
Pack-years	25.8±25	25.2±24	31.8±27	42.1±15	NA	NA	NA	NA
Reader Analysis	YES	YES	YES	YES	NA	YES	NA	YES
Kilovoltage Mean, Range	118 kVp, 80-140 kVp	120 kVp, 120-120 kVp	120 kVp, 120-120 kVp	120 kVp, 120-120 kVp	122 kVp, 120-140kVp	120 kVp, 100-140 kVp	120 kVp, 120-120 kVp	NA
Current Mean, Range	413 mAs, 36-795 mAs	263 mAs, 160-392 mAs	251 mAs, 160-386 mAs	65 mAs, 40-80 mAs	410 mAs 240-500 mAs	413 mAs, 152-425 mAs	200 mAs, 200-200 mAs	NA
Slice Thickness Mean, Range	3.30 mm, 1.0-6.0 mm	0.65 mm, 0.6-0.9 mm	0.70 mm, 0.6-0.8 mm	0.75 mm, 0.6-1.3 mm	1.0 mm, 1.0-1.0 mm	2.5mm, 1.5-3.0 mm	0.75mm, 0.75-0.75mm	NA

Definition of abbreviations: UIHC – University of Iowa Hospitals and Clinics; COPDGene – Genetics and Epidemiology of Chronic Obstructive Pulmonary Disease Study; INHALE – The Genetics Epidemiology of Lung Cancer Study; NLST – National Lung Screening Trial; LungX – International Society for Optics and Photonics LungX Challenge; LIDC – Lung Imaging Database Consortium; FDA – Food and Drug Administration Lungman Phantom; DDSM – Digital Database of Screening Mammography; M – malignant; B – benign; NA – not available/acquired; Apollo – analysis performed using Vida Apollo software; X – female sex; Y – male sex

The University of Iowa Hospitals and Clinics (UIHC) utilizes an electronic medical record system (Epic Systems, Verona, WI) for the collection and storage of patient data. With institutional review board (IRB) approval, the radiology reports of chest CT scans from 2008-2014 were text searched for the terms 'pulmonary nodule' or 'lung nodule' to identify potential subjects. The medical records of these potential subjects were further manually mined for eligibility inclusion. This dataset was previously reported upon in Dilger, 2016[37]. Subsets of the clinical dataset collected were used in **Chapter 3: Mathematical Prediction Models**, **Chapter 4: Segmentation**, and **Chapter 6: Application of QIC-RATE to Histoplasmosis Classification** as described below.

2.1.1. Longitudinal Cohort

In current clinical practice, nodules are often imaged more than once prior to diagnosis, as a proof of concept study, we identified a subset of clinical subjects which were imaged multiple times (2-6) prior to diagnosis of the nodule. A cohort of 30 subjects with a total of 92 clinical CT scans were used to assess the improvement of published post-imaging mathematical prediction models longitudinally described in **Chapter 3: Mathematical Prediction Models**.

2.1.2. Segmentation Cohort

Automated and semi-automated tools to improve the analysis workflow are of great need, particularly as they reduce the amount of required human interaction – thereby potentially reducing time,

effort, and subjectivity. It is likely that heterogeneous scanning parameters affect these automated tools in a manner that do not affect human readers. Due to the wide acquisition parameters and clinical scan representation of this dataset, we utilized a subset of four clinical cases for the development and testing of the semi-automated segmentation tools described in **Chapter 4: Segmentation.**

2.1.3. Histoplasmosis – NSCLC Cohort

Histoplasmosis is a fungal infection that is endemic to the Ohio River Valley area which often presents in imaging as a solitary pulmonary nodule resulting in a clinical diagnostic issue[38]. This presents a difficult clinical challenge, with many cases of histoplasmosis undergoing invasive procedures for nodule diagnosis. With this dataset, we utilized a case-control matched cohort of 71 clinical histoplasmosis controls and non-small cell lung cancer cases collected at the University of Iowa to assess the developed approach's robustness to difficult, potentially heterogenous data and compare the performance to observers as described in **Chapter 6: Application of QIC-RATE to Histoplasmosis**.

2.2. Multicenter Study Datasets

Large multicenter studies have exploited the benefits of CT applied to the lungs to investigate aims including the mortality reduction of early detected lung cancer[39] and characterization of chronic obstructive pulmonary disease (COPD)[40-42]. These prospective trials have standardized CT acquisition protocol to support the ability to extract high-quality, reproducible quantitative measures[41,42]. While pulmonary nodule detection was not the primary goal of some of these multicenter studies, the large populations of current and former smokers, enrolled in these multicenter trials made 'incidental' pulmonary nodule discovery a likely event for a subset of subjects. With proper database management, the pulmonary nodules discovered during research study imaging were tracked and, with subject consent, the clinical follow-up of a nodule were recorded. For this dissertation, we have exploited pulmonary nodule data collection from three of these studies to procure a high-resolution research cohort with pulmonary nodules. The following sections detail the purpose of the origin of each dataset and indicate the eligibility for diagnostic truth for inclusion in our dataset. Sample slice images from these datasets are shown in **Figure 2.1**.

2.2.1. National Lung Screening Trial

The NLST was a multicenter study to compare lung screening with LDCT and chest radiography in a subpopulation of individuals with a high risk of lung cancer[39]. Each NLST participant underwent a baseline and two annual screenings (total of three imaging time-points) using either LDCT or chest radiography. As the NLST's target endpoint was the detection of lung cancer, diagnostic follow-up tracking of detected pulmonary nodules was part of the study design.

The UIHC was one of the centers performing LDCT screening for the NSLT. As the raw CT data was available to be reconstructed at a smaller slice thickness, some of these subjects with histopathology proven diagnosis were included in our high-resolution research cohort. We performed visual confirmation of lobar location of histopathology diagnosed pulmonary nodule with follow-up information from NLST. Subjects from the NLST were included in the following chapters: **Chapter 3: Mathematical Prediction Models**, **Chapter 4: Segmentation**, and **Chapter 5: QIC-RATE**.

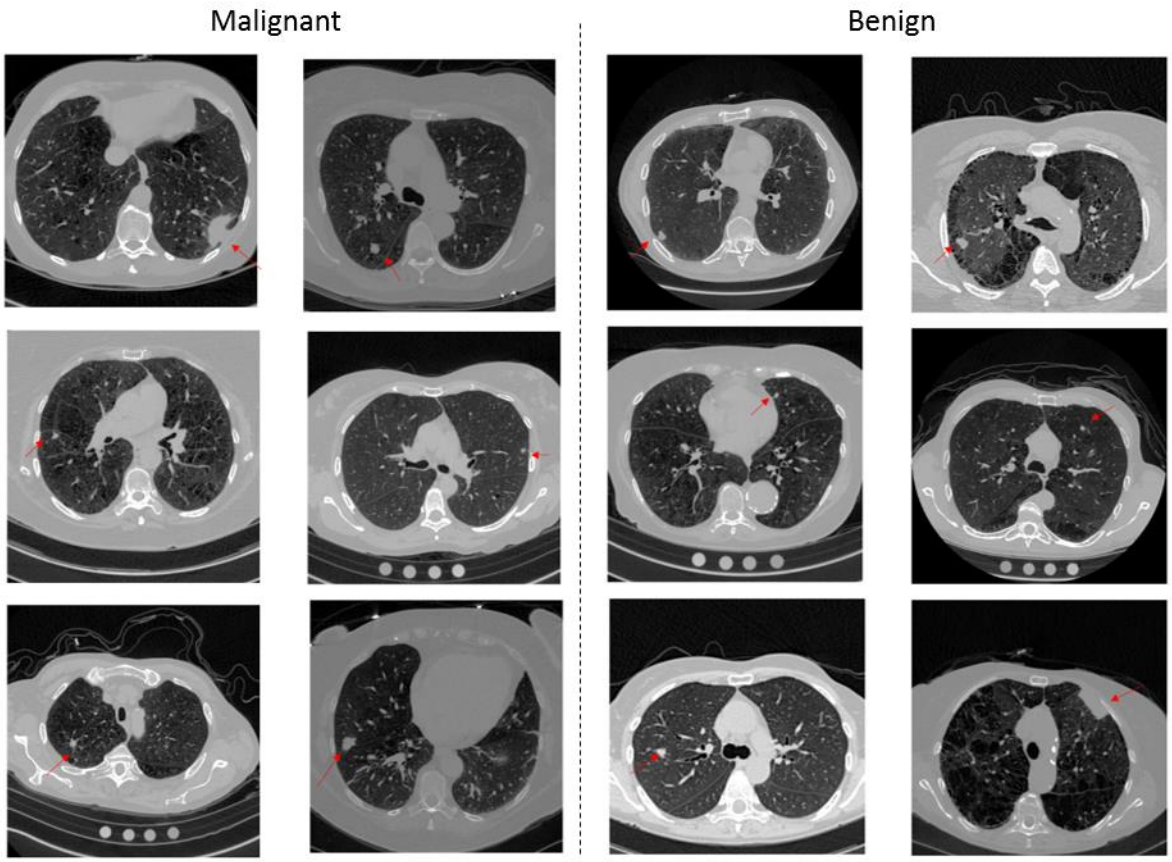

Figure 2.1: Sample slices from six malignant and six benign cases taken from the multicenter study datasets. Image screenshots produced in MatLab. Red arrows indicate nodule location.

2.2.2. Genetics and Epidemiology of Chronic Obstructive Pulmonary Disease

The Genetics and Epidemiology of Chronic Obstructive Pulmonary Disease (COPDGene) Study was a multicenter trial investigating the genetic and environmental exposures for COPD[40]. This study recruited subjects with various COPD Gold Stage (based on pulmonary function testing) including normal healthy subjects without COPD. Follow-up of imaging-identified lung nodules was not the primary aim of the COPDGene study; however, through ancillary funding, COPDGene began collecting nodule diagnostic information. We performed visual confirmation of lobar locations of histopathology diagnosed pulmonary nodules and evaluated follow-up scans for imaging-derived benign diagnosis of nodule resolution/stability. Subjects from the COPDGene study were included in the following chapters:

Chapter 3: Mathematical Prediction Models, **Chapter 4: Segmentation**, **Chapter 5: QIC-RATE,** and **Chapter 7: Application of QIC-RATE to Global Lung Measures**.

2.2.3. The Genetic Epidemiology of Lung Cancer Study

The Genetic Epidemiology of Lung Cancer (INHALE) study sought to evaluate the role of genes and the environment in the etiology of lung cancer, particularly with respect to racial susceptibility in the link between lung cancer and COPD[41]. Tracking nodule outcome (diagnosis, treatment etc.) was an integral part of the parent study. Subjects underwent non-contrast chest LDCT scans at both full inspiration and full expiration under a standardized protocol. Through a collaboration with Dr. Ann Schwartz, we have included a subset of subjects with confirmed diagnosis whose CT scans, demographics, and clinician information was collected. Subjects from the INHALE Study were included in the following chapters: **Chapter 3: Mathematical Prediction Models**, **Chapter 4: Segmentation, Chapter 5: QIC-RATE**, and **Chapter 7: Application of QIC-RATE to Global Lung Measures**.

2.3. Publicly Available Datasets

The datasets described in the previous two sections were acquired either through UIHC or through scientific collaborations at other institutions. Recently, efforts in the imaging community have been made to compile large, public datasets for researchers to have resources to build and to test methods of image processing and image machine learning. There has also been a greater emphasis on data transparency and sharing which has prompted the creation of centralized sources for finding publicly available imaging datasets. One which emphasizes imaging data collected of tumors is The Cancer Imaging Archive (TCIA) (http://www.cancerimagingarchive.net/)[43]. The dataset curation for these collections can be diverse and include information such as clinical factors, treatment factors, genetic factors, image segmentations, and follow-up data. In this dissertation, portions of several of these collections from TCIA were used to supplement questions/tests and validate pipeline.

2.3.1. Lung Imaging Database Consortium

The Lung Imaging Database Consortium (LIDC) is a collection of CT images that have been assessed by four radiologists[44,45]. This dataset was retrospectively collected from seven institutions and includes 1018 thoracic CT scans with/without lung nodules. As these were collected from clinical conditions, it included heterogeneous scanning parameters. Each scan was read by four radiologists who, if a they encountered a nodule $\geq$ 3mm in maximum in-plane diameter, used a computer interface to segment the nodule. They were also prompted to provide an assumed malignancy risk score (1, low-risk, to 5, high-risk). This data does not include histopathological diagnosis on the majority of nodules. Here, we have used the LIDC-IDRI in **Chapter 4: Segmentation** to assess the developed nodule segmentation tools.

2.3.2. Lungman Phantom for Segmentation

The Quantitative Imaging Biomarker Alliance (QIBA) challenge used a phantom with multiple nodules aimed at determining the repeatability of segmentation techniques and establishing standards for acceptable deviation among tools[46]. CT data was collected using the Food and Drug Administration Lungman phantom and consisted of repeated scans of the Lungman phantom with various layouts of synthetic nodules performed on a common scanner. For this dissertation document, we utilized the QIBA challenge dataset to further confirm segmentation results in **Chapter 4: Segmentation**.

2.3.3. International Society for Optics and Photonics LungX Challenge

The International Society for Optics and Photonics LungX Challenge was implemented in 2015 to compare performance of radiomic risk scores from participants using a common cohort of nodules[47,48]. It includes 72 diagnosed nodules retrospectively collected from the University of Chicago with IRB approval. All scans had been acquired on a common scanner model with consistent reconstruction parameters. Diagnostic assessment was confirmed through either pathology (malignant, benign) or imaging (benign). The LungX Challenge data was included in **Chapter 5: QIC-RATE**.

2.3.4. Curated Breast Imaging Subset of the Digital Database for Screening Mammography

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is a collection of breast mammography screening exams that was publicly released with histopathology proven diagnosis of either 'malignant' or 'benign'[49]. The CBIS-DDSM contains 2,620 scans with identified masses (tumors) and calcifications. For this dissertation, only a subset of scans identified by solid tumor masses were included for analysis. This dataset was used in **Chapter 8: Application of QIC-RATE to Breast Tumor Classification** to assess the transferability to other imaging dimensionalities and disease locations and to provide a more data-corrected comparison to other approaches.

CHAPTER 3: MATEHMATICAL PREDICTION MODELS

This chapter is adapted from "Post-Test Pulmonary Nodule Mathematical Prediction Models: Are They Clinically Relevant?" accepted for publication in European Radiology[50].

3.1. Significance and Background

Lung cancer is the leading cause of cancer-related deaths in the United States[4]. CT imaging is used to detect and to characterize lung nodules. Size-based guidelines exist to provide clinicians with criteria to assess the potential malignancy of pulmonary nodules including the Lung-RADs Assessment Categories, American College of Chest Physicians Clinical Practice Guidelines, and Fleischner Society Follow-Up Guidelines[33,51,52]. However, based on size alone, these have the potential to misclassify both small malignant nodules and large benign nodules leading to suboptimal treatment plans[53-55]. This is particularly true of first encounters, or 'de novo' nodules, which often fall into CT surveillance recommendations without access to growth information.

Pre-imaging lung cancer risk models have been produced which seek to stratify the individual's benefit from screening thereby reducing unnecessary radiation exposure on subjects with limited benefit from CT imaging[56]. To better characterize imaging-detected nodules, post-imaging mathematical prediction models (MPMs) have been developed using multivariate logistic regression models of known lung cancer risk factors such as family history, demographics, and radiologist-defined imaging characteristics to provide a malignancy risk stratification after an imaging encounter[57-61]. Previously, MPMs have been utilized on an ad-hoc basis by clinicians seeking standardized input from evidence-based models. However, recently, an MPM was incorporated into the British Thoracic Society's (BTS) Guidelines for Nodule Follow-up following an initial size-based stratification of risk (grade C recommendation) indicating a growing interest in the increased use of MPMs for day-to-day management of pulmonary nodule subjects[62].

This chapter compares four previously published post-imaging MPMs: the Mayo Clinic model (MC)[57], the U.S. Department of Veterans Affairs model (VA)[58], the Peking University model (PU)[60], and the Brock University model (BU)[59], on a large cohort of trial subjects and a longitudinal cohort of retrospective clinical subjects. As these MPMs were developed using different imaging parameters (clinical chest radiographs[57,60], clinical CT scans[58,60], or lung cancer screening CT scans[59]), different proportions of malignant cases (MC: 35%; VA: 54%; PU: 61%; BU 6%), and variable size distributions (mean size malignant/benign; MC: 17.8mm/11.6mm; VA: 18.9mm/14.8mm; PU: 21.3mm/17.2mm; BU: 15.7mm/4.1mm), we expect significant cohort dependence to be seen when each MPM was applied to an independent dataset.

While several studies have attempted to compare the accuracy of various post-imaging MPMs, they have reported performance (sensitivity, specificity) based on optimized threshold points for their unique study cohorts as opposed to the recommended thresholds associated with a given MPM[63-65]. These studies reported that independent cohort-optimized thresholds can vary greatly from the MPM thresholds and adjustments to the cut-off used affects sensitivity and specificity values[65]. This presents a lack of clarity in the appropriate cut-off point for a given MPM to be applied in the clinical context[66]. Here, we evaluate the current clinical usefulness of MPMs using the recommended thresholds and compare the performance to our study-optimized cut-offs.

3.2. Materials and Methods

3.2.1. Study Cohorts

As mentioned, the MPMs investigated here have been built and tested in diverse datasets including lung screening and incidental clinical cases. For this study, two cohorts of subjects with pulmonary nodules were investigated: a research cohort and a longitudinal clinical cohort. (**Table 3.1**).

3.2.1.1. Research Trial Cohort

The research cohort consisted of 317 subjects (80 malignant, 237 benign) included from two separate trials collecting high-resolution CT scans (217 COPDGene[40], 100 INHALE[41]). While neither study was specifically aligned with the recommendations for screening for lung cancer, both had de-novo nodules encountered during the course of imaging. Demographic and historical information was collected from participants in these trials and radiologist reports were generated to include descriptions of nodule findings. Further information about these studies is included in **Chapter 2**.

3.2.1.2. Longitudinal Clinical Cohort

The longitudinal clinical cohort was included in this study as a proof of concept on MPM stability over time and repeated imaging. The cohort consisted of 30 subjects (16 malignant, 14 benign) with 92 clinical CT scans retrospectively collected from the University of Iowa Hospitals and Clinics (**Table 3.1**). Further information regarding the origin of the clinical data is included in **Chapter 2.** For this assessment, we compared the performance of the MPM predictions at (a) the initial (incidental) imaging encounter on which the pulmonary nodule was identified (TP_I), (b) the final imaging encounter before diagnosis (TP_F), and (c) across all the imaging encounters between detection and diagnosis.

Table 3.1: Subject and nodule demographics of study cohorts.

Cohort	Demographics		Malignant	Benign
Research	Number of Subjects		80	237
	Age (years) (Mean, Range)		64.0 (41-87)	62.1 (40-86)
	Sex		54F : 26M	113F : 124M
	Pack-years (Mean, Range)		40.7 (0-80)	15.8 (0-50)
	Nodule Size (Mean, Range)		16.3mm (4-30mm)	9.2mm (4-30mm)
	LDCT screening eligible * (Yes: No)		48:32	85:152
	Lung-RADs category	2	5	59
		3	6	56
		4A	27	92
		4B	42	30
Longitudinal Clinical	Subjects		16	14
	Age (years) (Mean, Range)		46.5. (23-64)	61.1 (40-74)
	Sex		9F : 7M	10F : 4M
	Pack-years (Mean, Range)		21.2 (0-50)	14.2 (0-25)
	Nodule Size (Mean, Range)		18.9mm (3-48mm)	13.3mm (3-29mm)

Definition of abbreviations: F = female, M = male, LDCT = low-dose computed tomography; * : LDCT screening eligibility criteria based on age between 55 and 80, and smoking pack-years ≥ 30 pack

3.2.2. Mathematical Prediction Models

Four MPMs were assessed (MC[57], VA[58], PU[60], BU[59]); detailed descriptions of the MPM-specific equations and variable descriptions is provided in the **Appendix B.1**. Pertinent risk variables were manually extracted from subject records and a risk score from each MPM was calculated for each subject (**Table 3.2**). Unless the radiological report specifically indicated the presence of calcification, spiculation, or the absence of a border, nodules were considered non-calcified, non-spiculated, and smooth-bordered.

3.2.3. Statistical Analysis

Detailed information on the specific performance measures is included in the **Appendix A**. In brief, MPM raw prediction performance was assessed using area-under-the-curve of the receiver-operator characteristic (AUC-ROC) (Delong) and precision recall (AUC-PR) techniques. The Youden's J statistic was used as the calibrated threshold to produce sensitivity and specificity. The stability of the Youden thresholds was assessed using median absolute deviation (MAD) below 0.05 on sub-set sizes between 50 and 250 subjects using 41,000 naïve bootstrapping trials sampling without replacement; for additional details, refer to **Appendix B.2**. McNemar's tests was used for statistical difference between binary

classifications (inter-MPM and intra-MPM). For MPM-recommended thresholds, we assessed the performance by recommendation-induced misclassification of nodule or delay in ground-truth diagnosis. MPM-recommended categories were binarized into benign-tagged ('low-risk' or 'watchful waiting' and malignant-tagged ('high-risk' or recommended immediate additional work-up).

Table 3.2: Tabular form of mathematical prediction model's (MPMs) base equations. Risk variables are categorized into demographical (subject reported) and radiological (clinician reported) factors.

Risk variable		Units	MPM Coefficient			
			MC	VA	PU	BU
Demographical	Age	Years	0.0391	0.0779	0.07	0.0287
	Sex	F/M				0.6011
	Ever Smoker	Y/N	0.7917	2.061		
	Time of smoking cessation	Years		0.0567		
	Cancer history	Y/N	1.3388			
	Family history of cancer	Y/N			1.267	
	Family history of lung cancer	Y/N				0.2961
Radiological	Emphysema	Y/N				0.2953
	Upper lobe	Y/N	0.7838			0.6581
	Diameter[a]	MM	0.1274	0.112	0.0676	-5.3854*
	Spiculation	Y/N	1.0407		0.736	0.7729
	Smooth Border	Y/N			-1.408	
	Calcification	Y/N			-1.615	
	Nodule type	Solid : Y/N				0
		Part Solid: Y/N				0.377
		Non-Solid: Y/N				-0.1276
	Nodule count	Count				-0.0824
		Base Intercept/Offset	-6.872	-8.404	-4.496	0.2761

Note: Units are coded in clinical terms; for use in the equation(s), sex (F=1,M=0) and presence (Y=1,N=0) are numerically coded. To obtain a prediction value for a given MPM, multiply each coefficient by the subject's risk variable value and take the summation with the base intercept/offset. The resulting number is the x in the logistic equation: e^x/((1 + e^x)) = risk prediction. For example, performing the VA MPM prediction for a 62-year-old, never-smoker, with a 10mm nodule would yield x = (62*0.0779 + 0*2.061 + 0*0.0567 + 10*0.112 - 8.404) = -2.454; risk prediction from the logistic equation would yield, 0.079.
Definition of abbreviations: MPM – mathematical prediction model; MC – Mayo Clinic; VA – Veteran's Affairs; BU – Brock University; PU – Peking University; F – Female; M – Male; Y – Presence; N – Absence; *: In the BU model, nodule size is defined by (diameter in millimeters/10)^-0.5

3.2.4. Development of Application to Calibrate and Assess Local Datasets

Prior literature has noted the need for calibration on individual MPMs, but no easy tool exists for clinicians to both (a) determine which MPM is suitable for their population and (b) calibrate the MPM to their local population[57]. Some of these MPMs have an online calculator tool that can provide the risk assessment from the original calibration on a single patient[67-69]. However, these tools do not provide easy means for calibration on a large cohort.

As part of this study, we developed an open-source web-based application capable of performing cohort calibration and performance analysis measures discussed in this chapter. The application can be used to guide researchers and clinicians through the process of (1) creating a local dataset, (2) calibrating the MPMs to the local dataset, and (3) determining which of the four MPMs is the best fit for their population. The app was developed using R and the Shiny library for web-based application development. R is an open source programming language and software environment, originally developed for statistical computing and graphics, which has a wide range of contributed packages, similar to libraries, expanding functionality of the base code. Shiny is an R package which was developed to make it easy to build interactive web applications over existing R code[70,71].

3.3. Results

3.3.1. Calibrated Thresholds Equalize Performance Among MPMs

The four models (MC, VA, PU, BU) were applied to the research cohort (N = 317, 80 malignant, 237 benign) yielding four risk scores (one per MPM) per subject which were compared with the nodule's known diagnosis (**Figure 3.1**, solid line). The impact of risk stratification based on the calibrated threshold (**Table 3.3**, **Table 3.5**) and MPM-associated categories (**Table 3.6**) were applied to the predictions (**Figure 3.1**, solid line). The optimal AUC-cutoff (**Figure 3.1**, dashed line) was derived for each of the models. The MC (AUC: 0.63) and BU (AUC: 0.61) MPMs achieved the best separation between classes on this cohort compared to PU (AUC: 0.55) and VA (AUC: 0.51) MPMs. The MC and BU MPMs were both statistically significantly better than the VA MPM ($p = 0.02$); all other pairwise comparisons of significance yielded p-values above the assigned alpha (0.05). No MPM significantly outperformed all others, revealing relative similarity in their calibrated discriminatory capability between malignant and benign nodules.

Table 3.3: Performance measures using cohort-derived optimized thresholds using Youden's J Statistic (Figure 1, dashed lines).

	Nodule Size	MC	VA	PU	BU	BU-BTS
AUC-ROC	**ALL**	**0.80**	**0.74**	**0.75**	**0.81**	**0.78**
	<6mm	0.67	0.69	0.58	0.64	NA
	≥6mm to < 8mm	0.48	0.52	0.60	0.56	NA
	≥ 8mm to < 15mm	0.74	0.66	0.70	0.72	0.72
	≥ 15mm	0.69	0.57	0.66	0.69	0.69
AUC-PR	**ALL**	**0.63**	**0.51**	**0.61**	**0.55**	**0.66**
	<6mm	0.13	0.11	0.33	0.11	NA
	≥6mm to < 8mm	0.11	0.16	0.11	0.14	NA
	≥ 8mm to < 15mm	0.40	0.31	0.39	0.47	0.47
	≥ 15mm	0.79	0.68	0.75	0.75	0.75

Definition of abbreviations: AUC-ROC area-under-curve of receiver-operator characteristic; AUC-PR – area-under-curve of precision-recall; MC – Mayo Clinic; VA – Veteran's Affairs; PU – Peking University; BU – Brock University; BTS – British Thoracic Society

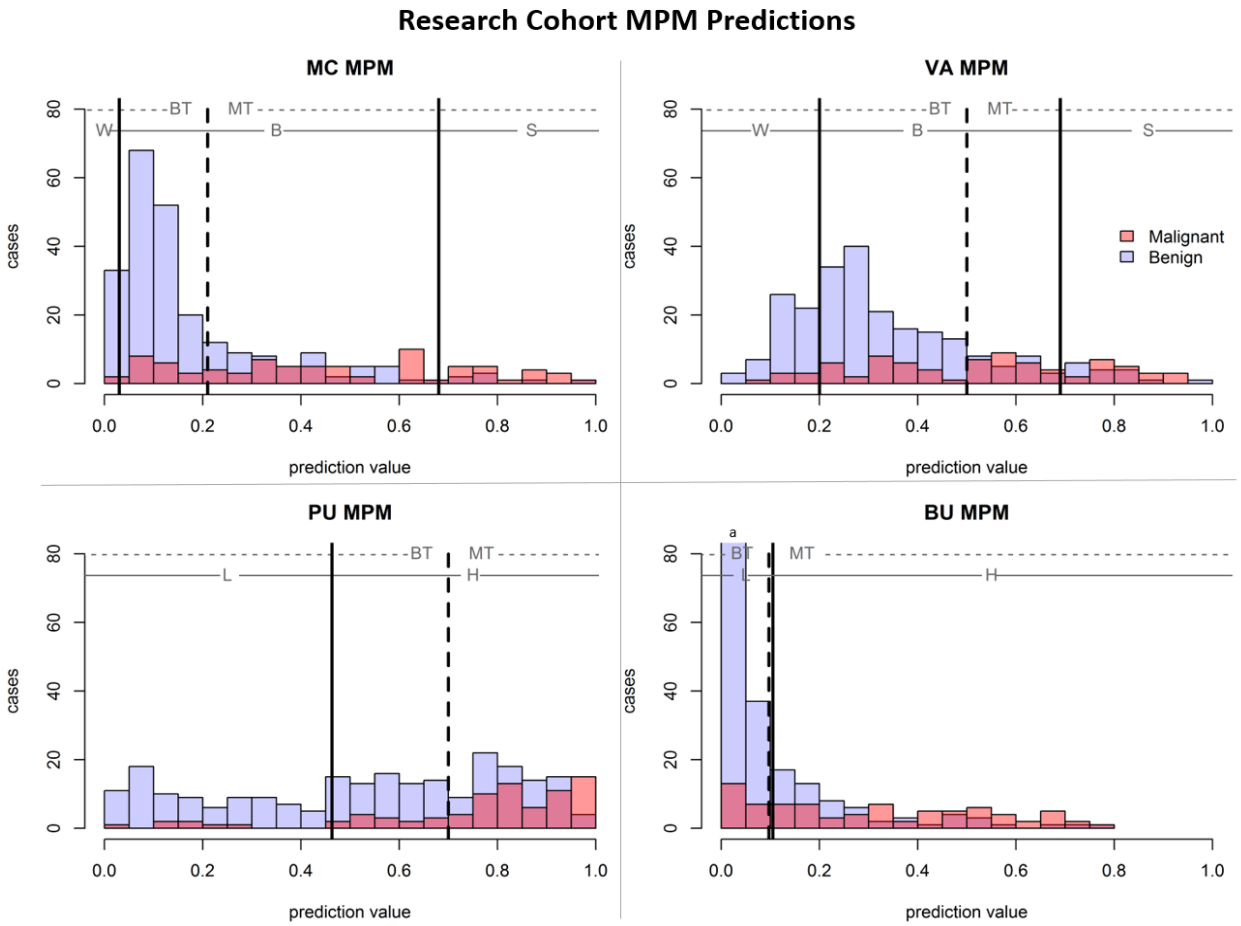

Figure 3.1: Histograms of MPM predictions split based on true nodule classification. Solid lines indicate MPM-derived thresholds with MPM-assigned categories of watchful-waiting (W), biopsy (B), surgery (S), low-risk (L), or high-risk (H). The dashed line indicates

3.3.2. Youden Threshold Stability and Application development

A naïve bootstrapping without replacement method using 10,000 combinations was performed on the research cohort using set sizes between N = 50 and N = 250 by increments of 5 subjects. The median absolute deviation (MAD) in Youden threshold was calculated for each set size, N. MAD is robust to outliers and provides an unsigned (absolute) measure of deviation of the set size's Youden threshold which is blinded to the full cohort's Youden threshold. Youden threshold stability was determined when the set MAD was below 0.05 (Figure A.1). The set median Youden threshold was also compared to the Youden threshold of the full cohort (N = 317) (**Figure 3.2**). Viewed together, this demonstrates the stability, blinded to the full cohort, of the Youden and that the stability converges on the full cohort's Youden. Testing the Youden threshold stability (MAD < 0.05) at different calibration set sizes demonstrated stability at 100 subjects for three MPMS (MC, BU, PU) and stability at 145 subjects for all four MPMs, thus establishing a recommended clinical calibration dataset size. For additional details refer to **Appendix B.2.**

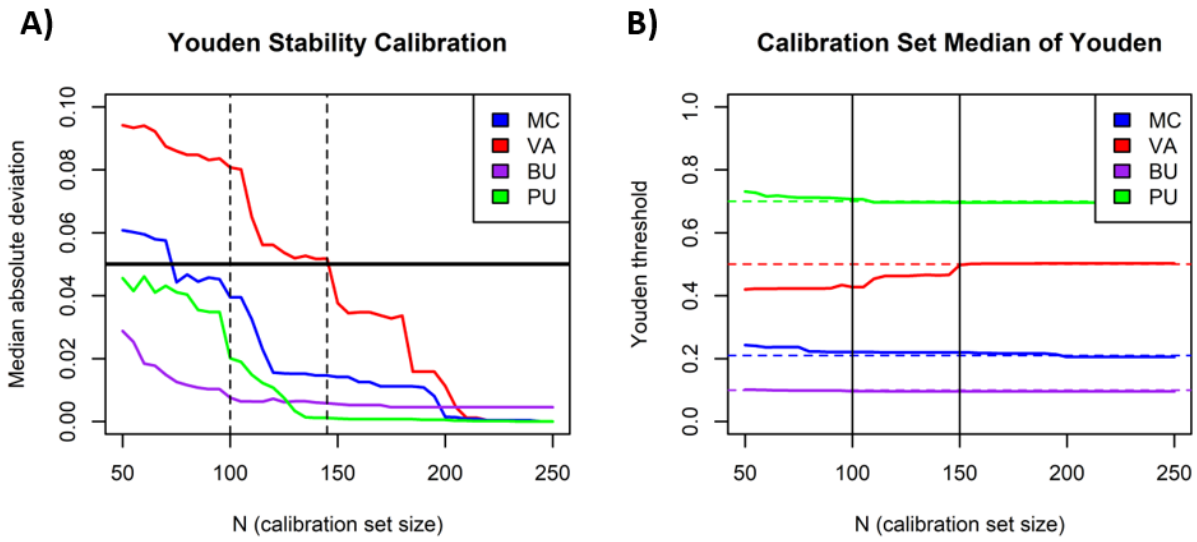

Figure 3.2: Youden threshold stability compared to calibration set size. A) Stability was considered met at a level of 0.05 in median absolute deviation from trial median Youden. B) Convergence of trial median Youden (dashed lines) to full cohort Youden (solid lines).

The developed app is hosted at https://www.i-clic.uihc.uiowa.edu/resources/sieren/mpm/) and the origin code is freely available through GitHub (https://github.uiowa.edu/APPIL/MPM) under the Creative Commons license Attribution-Share-Alike. **Figure 3.3** shows a flowchart of how the application can be used to perform local population calibration. In brief, the user simply requires a spreadsheet containing the relevant demographic and imaging variables for a representative sample of their patient population. Calibration is then performed using the statistical tests described previously, with the ability to provide thresholds that are customizable to desired level of sensitivity and specificity.

3.3.3. Calibrated Thresholds Out-perform the Original Recommended Thresholds in Work-up Categorization

Using the MPM-associated categories, up to 25% of the malignant tumors would have been assigned low-risk, while 25.3% to 97.5% of benign tumors would have been recommended for further work-up. The BU MPM was the only model to have agreement between the Youden-optimized calibrated threshold (0.10) and the MPM-associated guidelines (0.10) for the full cohort; however, in nodules ≥ 15mm the Youden optimized threshold was much higher (0.32). Furthermore, McNemar's comparison between the optimal and recommended thresholds demonstrated significant difference between the classification accuracy of three of the MPMs (MC, VA, PU) with $p < 0.001$, indicating that calibration to the local dataset improves discriminative prowess over original MPM-associated risk categorizations. As the BU Youden optimal threshold was nearly identical to the recommended, there was no statistical significance $p=0.99$, this stability indicates the BU MPM-associated thresholds were already well calibrated for this cohort.

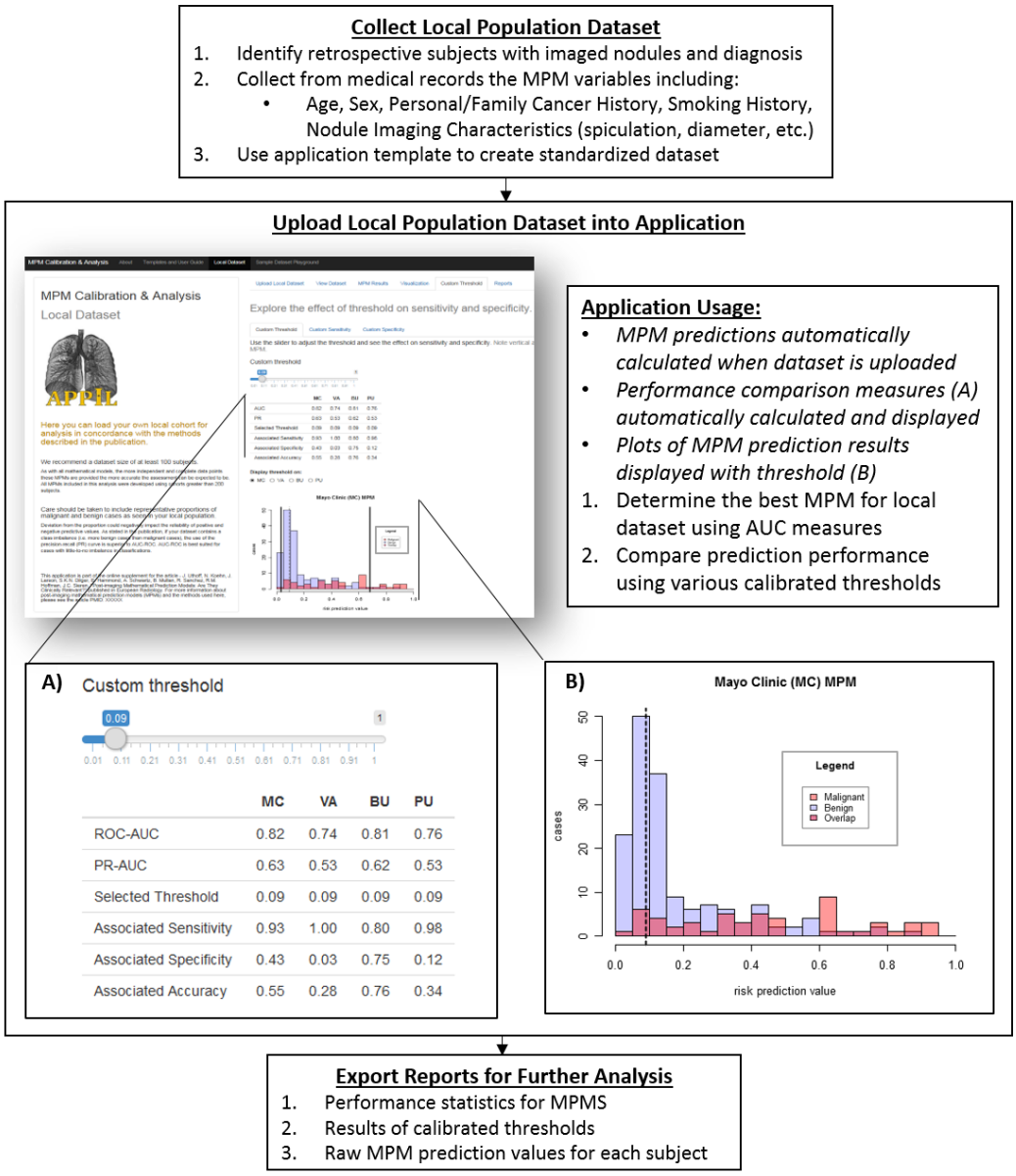

Figure 3.3: Flowchart of developed application to determine which MPM and calibration threshold to use for a local population. Definition of abbreviations; MPM – mathematical prediction model

3.3.4. Comparison to Fleischner Size-based Clinical Management Recommendations

The Fleischner Guidelines for Management of Incidental Pulmonary Nodules Detected on CT indicates that solid pulmonary nodules have a differential follow-up using three size-based thresholds (<6mm; 6-8mm; >8mm)[51]. To compare the Fleischner to the calibrated MPMs, the size-threshold of ≥8mm was used for 'high-risk' prediction and <8mm for 'low-risk' prediction. **Table 3.4** shows the breakdown for these categories and the clinical consequences of the follow-up recommendations. McNemar's analysis demonstrated that the Youden calibrated predictions for all four MPMs was statistically superior ($p < 0.01$) than the Fleischner designations.

Table 3.4: Distribution and potential clinical issues of Fleischner Follow-up Guidelines on the Research Cohort.

Category	Malignant	Clinical issue	Benign	Clinical issue
< 6mm	5	5 malignant wait indefinite amount of time (No routine follow-up)	59	No clinical issue since no follow-up
≥ 6mm to < 8mm	6	6 malignant wait 6-12 months	56	56 benign have extra CT in 6-12 months (increased radiation)
≥ 8mm	69	69 malignant wait 3 months, PET or Biopsy	122	122 benign have extra CT in 3 months

3.3.5. Calibrated Thresholds Improve Specificity in Nodules ≥8mm

Size is a common variable among the MPMs and is prominent in current management guidelines. An accurate MPM risk assessment would be most clinically interesting and powerful on the nodules ≥8mm to <15mm at baseline with 5-15% probability of malignancy in Lung-RADS – in this study, 119 nodules (27 malignant, 92 benign). The best compromising MPM at this size category was the PU model, which using MPM-associated thresholds achieved 97% sensitivity but only 36% specificity; applying Youden optimal threshold achieved 67% sensitivity and improved specificity to 61% (**Tables 3.5-6**). Using the MPM-associated threshold, VA model would have only missed one malignant nodule, but at the cost of 79 benign nodules undergoing biopsy (75 cases) or surgery (4 cases); the optimized threshold improved VA MPM specificity for the nodules between 8-15mm to 82%. The MC model was the only MPM to completely reduce wait-time on malignant tumors sending 26 to biopsy and 1 to surgery; however, all benign tumors would have also been assigned to biopsy (91 cases) or surgery (1 case); here, applying optimized thresholds significantly improves specificity to 70% with sensitivity of 70%. In considering nodules between 8 and 15 mm in diameter, the MPM-associated recommendation thresholds for work-up have little benefit in tradeoff between sensitivity and specificity. Applying optimized thresholds improves specificity at the cost of some sensitivity.

3.3.6. Size-Exclusion Prior to MPM in BTS Guidelines Appropriate

The BU model is unique as it has been incorporated into the BTS guidelines for management of nodules; per BTS decision flowchart, only nodules ≥ 8mm are to be assessed with the BU MPM[31]. **Table 3.3** demonstrates the BU accuracy for that size-based subset. On our cohort, following the BTS exclusion of nodules < 8mm in diameter would have meant 11 malignant and 115 benign nodules would not be assessed with the BU due to size-exclusion. Applying the BU to the size-excluded, no malignant and 9 benign nodules are labeled 'high risk' by the BU MPM. Of the 11 malignant size-excluded nodules, one is recommended to be 'discharged', four are recommended for a 1-year follow-up CT, and six are recommended for a 3-month CT -indicating the need for more sophisticated discrimination techniques geared towards small nodules. The BTS recommendation to not include BU prediction on small nodules is appropriate, and as the BU threshold did not change with calibration, the recommended decision of 10% risk (0.1 prediction value) is well founded.

Table 3.5: MPM optimized categories size-breakdown of nodule risk prediction using Youden threshold.

MPM	Size	Malignancy probability and associated recommendation	
MC		< 21% Low Risk	≥ 21% High Risk
	All	**19M: 180B**	**61M: 57B**
		24% malignant wait	**24% benign immediate work-up**
	<6mm	5M:59B	0M:0B
		100% malignant wait	0% benign extra procedures
	≥6mm to < 8mm	5M:51B	1M:5B
		83% malignant wait	9% benign extra procedures
	≥ 8mm to < 15mm	8M: 67B	19M: 25B
		30% malignant wait	30% benign immediate work-up
	≥ 15mm	2M: 2B	40M: 28B
		5% malignant wait	93% benign immediate work-up
VA		< 50% Low Risk	≥ 50% High Risk
	All	**34M: 197B**	**46M: 40B**
		43% malignant wait	**17% benign immediate work-up**
	<6mm	5M:59B	0M:1B
		100% malignant wait	2% benign extra procedures
	≥6mm to < 8mm	6M:56B	0M:0B
		100% malignant wait	0% benign extra procedures
	≥ 8mm to < 15mm	16M: 75B	11M: 17B
		59% malignant wait	18% benign immediate work-up
	≥ 15mm	7M: 8B	35M: 22B
		17% malignant wait	73% benign immediate work-up
BU		< 10% Low Risk	≥ 10% High Risk
	All	**19M: 178B**	**61M: 59B**
		24% malignant wait	**25% benign extra procedures**
	<6mm	5M:59B	0M:0B
		100% malignant wait	0% benign extra procedures
	≥6mm to < 8mm	6M:55B	0M:1B
		100% malignant wait	2% benign extra procedures
	≥ 8mm to < 15mm	9M: 61B	18M: 31B
		33% malignant wait	34% benign immediate work-up
	≥ 15mm	0M: 1B	42M: 29B
		0% malignant wait	97% benign immediate work-up
PU		< 70% Low Risk	≥ 70% High Risk
	All	**18M: 154B**	**62M: 83B**
		22% malignant wait	**35% benign immediate work-up**
	<6mm	3M:54B	2M:5B
		60% malignant wait	9% benign extra procedures
	≥6mm to < 8mm	3M:34B	3M:22B
		50% malignant wait	39% benign extra procedures
	≥ 8mm to < 15mm	9M: 56B	18M: 36B
		33% malignant wait	39% benign immediate work-up
	≥ 15mm	6M: 18B	36M: 12B
		14% malignant wait	39% benign immediate work-up

Table 3.6: MPM-assigned categories size-breakdown of nodule risk prediction using published thresholds.

MPM	Size	Malignancy probability and associated recommendations		
MC		<3% Watchful waiting	3-68% Needle biopsy	>68% Surgery
	All	**0M:6B**	**61M:224B**	**19M: 7B**
		0% malignant wait	**97.5% benign extra procedures**	
	<6mm	0M:6B	5M:53B	0M:0B
		0% malignant wait	89.8% benign extra procedures	
	≥6mm to < 8mm	0M:0B	6M:56B	0M:0B
		0% malignant wait	100% benign extra procedures	
	≥ 8mm to < 15mm	0M:0B	26M:91B	1M:1B
		0% malignant wait	100% benign extra procedures	
	≥ 15mm	0M:0B	24M:24B	18M:6B
		0% malignant wait	100% benign extra procedures	
VA		<20% Watchful waiting	20-69% Needle biopsy	>69%Surgery
	All	**7M : 58B**	**51M: 163B**	**22M: 16B**
		8.8% malignant wait	**75.5% benign extra procedures**	
	<6mm	1M:29B	4M:30B	*0M:0B*
		20% malignant wait	50.9% benign extra procedures	
	≥6mm to < 8mm	2M:16B	4M:40B	*0M:0B*
		33% malignant wait	71.4% benign extra procedures	
	≥ 8mm to < 15mm	1M:13B	25M:75B	*1M:4B*
		3.7% malignant wait	85.9% benign extra procedures	
	≥ 15mm	3M:0B	18M:18B	*21M:12B*
		7.1% malignant waits	100% benign extra procedures	
BU		<10% Low risk	>10% High risk	
	All	**20M: 176B**	**60M:61B**	
		25.0% malignant wait	**25.3% benign extra procedures**	
	<6mm	5M:59B	0M:0B	
		100% malignant wait	0% benign extra procedures	
	≥6mm to < 8mm	6M:55B	0M:1B	
		100% malignant wait	1.8% benign extra procedures	
	≥ 8mm to < 15mm	9M:61B	18M:31B	
		33.3% malignant wait	33.7% benign extra procedures	
	≥ 15mm	0M:1B	42M:29B	
		0% malignant wait	96.7% benign extra procedures	
PU		<46.3% Nodule considered benign	>46.3% Nodule considered malignant	
	All	**8M: 87B**	**72M:150B**	
		10% malignant wait	**63.3% benign extra procedures**	
	<6mm	2M:28B	3M:31B	
		40% malignant wait	52.5% benign extra procedures	
	≥6mm to < 8mm	0M:22B	6M:34B	
		0% malignant wait	60.7% benign extra procedures	
	≥ 8mm to < 15mm	3M:33B	24M:59B	
		7.1% malignant wait	64.1% benign extra procedures	
	≥ 15mm	3M:4B	39M:26B	
		7.1% malignant wait	86.7% benign extra procedures	

3.3.7. Limited Benefit in MPM Tracking Longitudinally

We investigated the improvement in MPM performance over repeated imaging time-point on a clinical, longitudinal dataset of nodules imaged up to 6 times (average 3.1±1.1) prior to diagnosis (**Figure 3.4**). The average number of days between sequential patient imaging encounters was 214 days (±338 days) with malignant nodules tending to have a slightly longer time between scans (218 days ±368) compared to benign nodules (197 days±305).

The VA model was the only MPM to also decrease the percentage of benign nodules at TP_F that were categorized as high risk. The TP_I AUCs (MC: 0.62-0.96; VA: 0.65-0.96; BU: 0.51-0.90; PU: 0.70-0.98) were consistently higher than the TP_F AUCs in three of the MPMs (MC: 0.56-0.94; VA: 0.34-0.78; BU: 0.53-0.92; PU: 0.44-0.88). McNemar's p-value between TP_I and TP_F showed no statistical significance between MPM predictions at TP_I and TP_F (MC: 0.76, VA: 0.08, BU: 0.91, PU: 0.18), indicating no improvement to MPM risk assessment closer to diagnosis. This data suggests that MPM risk should not be incorporated into longitudinal evaluation of detected pulmonary nodules.

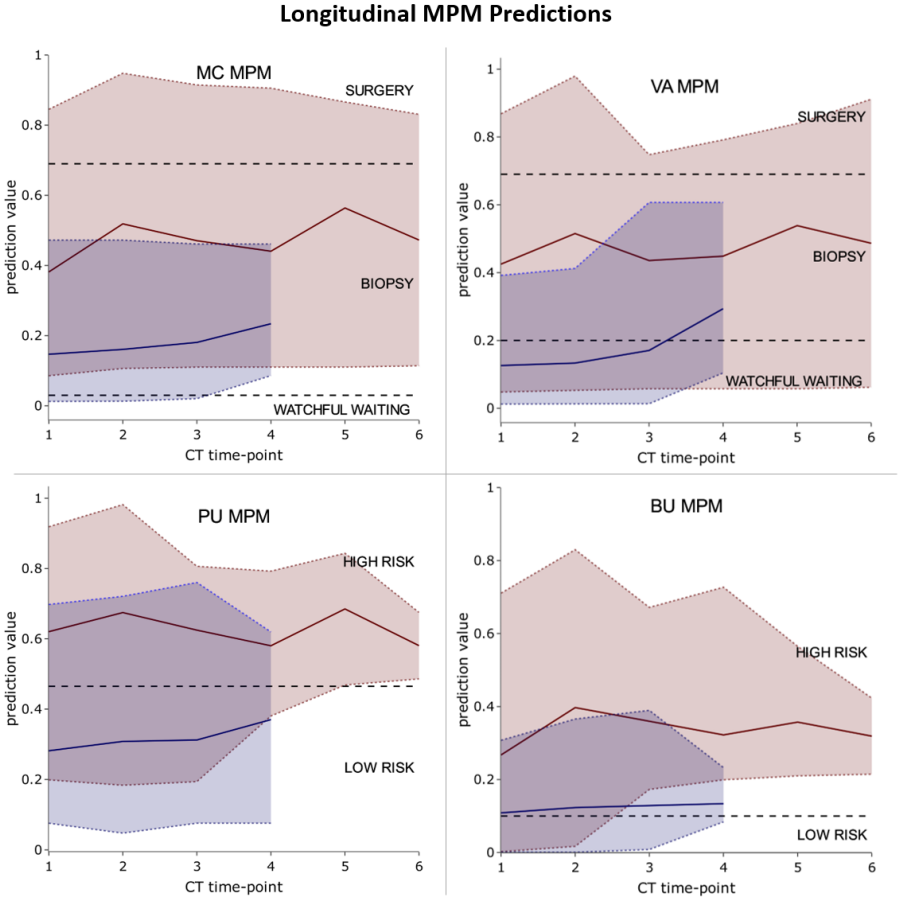

Figure 3.4: MPM prediction value over CT number on longitudinal cohort. The range in prediction values for malignant (red) and benign (blue) are shown with minimum and maximum values indicated by dashed colored lines. The average prediction value for the two classes is shown with the solid colored lines. Black dashed lines indicate Youden thresholds. Definition of abbreviations: MC – Mayo Clinic; VA – Veteran's Affairs; PU – Peking University; BU – Brock University; MPM – mathematical prediction model; CT – computed tomography

3.4. Discussion

We have applied four post-imaging MPMs to a large cohort of trial subjects and to a longitudinal cohort of clinical subjects. To our knowledge, this is the first study to compare MPMs by both the MPM-associated categories and AUC-derived (calibrated) classifications and to observe of MPM stability over longitudinal scans.

Recent alignment of size-based recommendations indicates that nodules ≥ 8mm in maximum diameter are at a heightened risk of malignancy[33,51,62]. Hammer et al. investigated eight risk calculators on a cohort of 86 nodules (59 malignant), showing a consistent under-estimation of malignancy risk. Here, we have a smaller proportion (25%) of malignancies in our cohort, yet our results concur with the assessment that care needs to be taken when assessing larger nodules (≥8mm) with these MPMs [63]. The applied BU model on the ≥8mm sub-cohort also demonstrated an under-estimation of true malignancy risk with an over-estimation of risk on benign nodules. Given average nodule size in the MPM development cohorts was larger than 8mm, it would be likely that the development-cohorts size bias would lead to more large benign nodules being tagged as suspicious.

Chung et al. recently validated the BU model on two large clinical cohorts showing that while the full model achieved AUCs of 0.901-0.911, the AUC-derived optimal threshold was 1.8-4% lower than the recommended BTS guidelines; this is a difference of 4-9% in sensitivity[72]. However, that study contained a significant size-bias between benign and malignant cases. While nodule diameter is not a variable in the BU model, the BTS flow-diagram applies the BU model only to nodules ≥8mm diameter (≥300mm^3 volume). Here we have applied the BU model in the manner recommended by BTS and demonstrated than all 11 below the size-stratified malignant nodules had a BU less than the threshold 10%. In practice, these malignant tumors would have remained untreated for at least 3 months before additional imaging.

While the BTS closely followed the original BU model study for this risk threshold, many independent surveys of MPMs have relied solely on the threshold derived from their cohort's AUC-ROC optimum[63-65]. Here we have displayed both the AUC-derived threshold from our cohort as well as the MPM-derived thresholds. When using our cohort-derived optimal cutoff point, MPM specificity was higher (65.0-83.0%) than through using the MPM-derived assigned categories (2.5%-74.7%), but MPM sensitivity was lower (58.0-78%) compared to MPM assigned categories (75.0%-100%). Based on MPM assigned categories, only the MC model would have detected 100% of malignancies at the imaging time point, but this is at the cost of requiring biopsy/surgery for all benign nodules. It is important to note that some studies have reported high AUCs of MPMs in their independent cohorts, but these studies have looked solely at the AUC-derived thresholds to assess MPM performance[64,66].

Our study has several limitations. First, the mean nodule size of the cohorts was smaller than those used to develop the MPMs. As nodule size was a common variable among the MC, VA, and PU

MPMs, this could have affected the prediction results. Second, the MPMs investigated here use subject-provided demographic/historical information and radiologist-described image characteristics, both of which can suffer from subjective variability and completeness. Radiologist variability is more easily investigated and has been shown to be different between radiologists as well as within a radiologist on so-called "coffee-break" reads in which a period of time is placed between repeated analysis[73,74]. While to a certain extent, the variability is built into the risk models in the development dataset, the modeling of noisy data is likely different between the development cohort and the user-end radiologist. Maiga et al. compared the MC model with clinician assigned risk from qualitative statements of cancer risk, showing that the current trend of qualitative risk statements for malignancy are highly variable and recommend a standardized scale for clinicians to follow[75]. Recent advances in CT including dose reduction techniques and reconstruction algorithms, have the potential to affect signal-to-noise ratio within the scan, thereby a potential source of variation that could affect both radiologist/reader efficiency and consistency. We do believe some of this variation is already contained within the development of the MPMs given the diverse (often clinical) datasets on which they were developed. Interesting to this point, the Mayo Clinic model (chest radiographs) performs on par with the Brock University model (low dose CT). Our cohort included only solid nodules, further studies are required to determine if MPM performance is affected when used on cohort of sub-solid tumors. Our research cohort consisted of 25.2% malignant cases and longitudinal clinical cohort 53.3% malignant cases; the MPMs compared here were developed on cohorts of subjects with difference malignancy rates (MC: 35%; VA: 54%; BU 6%; PU: 61%). We have included the AUC-PR measure to further describe the discrimination ability of MPMs in cohorts with disproportionate numbers of malignant and benign cases.

With the move towards digitized healthcare reporting and standardization of care, computer-based risk models have a natural place in the decision pipeline. There is a benefit to adding fully-automated, non-subjective systems with high performance to supplement radiologist reads with additional risk assessments. Efforts to develop tools which do not incur user subjectivity have been previously described; Mehta et al. compared the MC MPM with three multi-variate models developed with volumetric features extracted from semi-automatic (single click) segmentation of the nodule[65]. Machine learning for the assessment of lung cancer risk have been further developed to reduce extraction variability[10,13,15,21,22,25,76].

The number of lung nodules detected is set to increase with broad implementation of lung cancer screening programs . To make the screening and detection power of CT efficient and safe in practice, there is a great need for better informed decision making. Given proper assessment and application, post-imaging risk models have the potential to improve decision making processes. While standardization and wide-spread usage of these automated techniques has yet to happen, MPMs are being utilized in clinics

today. This paper has demonstrated the need for clarification in malignancy thresholds reported and demonstrated the cohort dependence built into these MPMs. We thereby recommend if an MPM is to be utilized for newly detected pulmonary nodules, that it is first calibrated with a retrospectively collected dataset (≥100 subjects) from the utilizing institution to ensure a locally optimal threshold value. We have developed an easy to use web-based application to assist institutions in performing MPM calibration and comparison of performance metrics between models. The application allows MPM discriminative power to be assessed using either AUC-ROC (balanced cohort) or AUC-PR (unbalanced cohort) measures and provides sensitivity and specificity. The lack of improvement in risk prediction from these MPMs over time suggests caution in the utility of these tools during surveillance stage of clinical management.

We have demonstrated that while MPM risk predictions are relatively stable across imaging time-points, there is a lack of evidence to their utility on an independent cohort without first performing cohort calibration. Based on our results, it is paramount that in the discussion and analysis of MPMs that the clinical applicability is thoroughly vetted by using the MPM-derived recommendations rather than the cohort-derived Youden thresholds as significant variations in performance ensue. Furthermore, we assessed this relevance to Lung-RADs associated 'indeterminate' or 'suspicious' nodules (between 8mm and 15mm in diameter), finding while MPM determination of malignancies in this category is improved over smaller nodules there is a stark increase in the number of recommended invasive procedures on benign tumors.

We have demonstrated the predictive capabilities of post-imaging MPMs developed with subject provided demographic/historical information and radiologist described imaging features. While predictive performance can be improved through calibration of the MPMs onto a locally representative dataset, there is still only moderate predictive capabilities, mostly centered around the size of the nodule. A weakness of these MPMS is the reliance on human provided features which can be sensitive to memory/extraneous circumstances (familial history knowledge) and reader subjectivity/variability; it also requires additional time to implement as human-described features need to be manually entered into the system. There is a need for additional predictive accuracy without need for additional human effort, which lends nicely into the use of automatically extracted quantitative imaging characteristics which can be supplied to machine learning algorithms to provide additional risk assessment. **Chapter 5: QIC-RATE** describes the method we developed for generating a risk assessment from these features. However, before these features can be automatically extracted from the imaging datasets, relevant regions/volumes of interest need to be identified to ensure feature relevance, the following **Chapter 4: Segmentation** details the assessment of semi-automated segmentation tools to streamline this process.

CHAPTER 4: SEGMENTATION

This chapter focuses on the segmentation results and discussion, methodology is explained in depth in **Appendix C**.

4.1. Introduction

The previous chapter demonstrated the potential utility of mathematical risk assessment tools for clinicians to make informed decisions regarding a patient's work-up. We concluded that while MPMs can be more clinically useful after calibration to a local cohort, improvement to the discriminatory ability in needed. Once source of limitation of MPMs come from the use of radiologists-described features which can be subjective and variable within and between readers. Also, patient-provided historical information which can be limited by knowledge (family history) and memory (smoking history with multiple cessation attempts). A method of risk assessment utilizing objective measures could provide greater performance and stability with little added human effort. One such method is automatic imaging feature extraction followed by machine learning for classification or clustering, these methods will be explored in **Chapter 5: QIC-RATE**.

Pre-processing for feature extraction requires image segmentation, which partitions the image into regions of interest (ROI). Manual volumetric segmentations by expert users are time consuming to generate and user subjectivity can influence both inter- and intra-observer agreement. Many algorithms for volumetric nodule segmentation have been developed and reported in the literature, from commercial, proprietary systems to research-driven academic tools; however, these segmentation methods are not widely available to researchers[46,77-79]. In a recent study from the Quantitative Imaging Network, of which we were a collaborating center, it was systematically shown that segmentation method can lead to differences in subsequent extraction of imaging features, including nodule volume[80]. This promotes the need for increased standardization in segmentations for quantitative analysis of disease. In this study, we compared the performance of easily built segmentation pipelines in easily accessible image-processing environments on a common test cohort. We explored if improvements in performance could be attained by developing an in-house graph-cuts based segmentation method. Our recommendations regarding lung nodule semi-automated segmentation approaches are based on the performance in the following areas: segmentation error, segmentation repeatability, ease of use, and customization.

4.2. Methods

4.2.1. Study Cohorts

This study included two cohorts of nodules aimed at assessing [1- Tool Flexibility] ability for segmentation method to be flexible to different protocols and parameters and [2 – Variability Accuracy]

testing segmentation method on nodules segmented by four radiologists. **Table 4.1** indicates demographical and scanning parameter ranges.

Table 4.1: Demographic and scanning parameters for segmentation tool comparison.

	24 Malignant	12 Benign
Age Mean	55 ± 7 years.	61± 10 years.
Sex (Female: Male)	10:14	8:4
Kilovoltage Range, Mean	80-140 kVp, 120 kVp	80-140 kVp, 115 kVp
Current Range, Mean	40-500 mA, 410 mA	40-600 mA, 410 mA
Slice Thickness Range, Mean	0.6-5.0mm, 2.75mm	0.6-5.0mm, 3.0mm
Nodule Size Range, Mean	6-29 mm, 15.5mm	8-24 mm, 16.2mm

4.2.1.1. Tool Flexibility Cohort

This dataset incorporated different diagnoses with 12 malignant and 12 benign confirmed with histopathology. Segmentation difficulty was determined by presence of extra-nodular attachments and irregular border or non-spherical shape (12 challenging, 12 fair). The overall demographics and scanning parameters of the subjects were diverse. Within this cohort, eight CT scans came from multi-center trials (4 NLST, 4 COPDGene) and 16 were retrospective clinical cases from the University of Iowa Hospitals and Clinics [26,40].

4.2.1.2. Variability Accuracy Cohort

To compare the tools against a 'more complete truth' the cohort included 12 scans from the LIDC with consensus tracings from four blinded radiologists' manual tracings used as the truth [45]. Solitary pulmonary nodules (<30mm in listed equivalent diameter), with confirmed primary lung cancer, DICOM CT data available with four radiologist annotation segmentations resulted in 24 cases, from which 12 were randomly selected. Annotations were extracted for each of the radiologists and a "consensus" tracing was generated using the Agreement Analysis Toolbox (Lampert, MatLab)[81]. The calculated agreement method used in this study was the level-set maximizing likelihood, LSML, maximizing the posteriori probability. These twelve scans were also used to assess the relative variability in segmentation compared to four experts.

4.2.2. Manual Segmentation Process

The main objective of a segmentation process is to produce a ROI; in this study, we explore the creation of a nodule ROI. An appropriate segmentation of the nodule often requires the removal of invalid

regions, such as vessels, chest wall, and pleural elements. As nodules can be directly interacting with an adjacent structure, segmentation requires careful consideration of object boundaries.

Using the Apollo software (Vida Diagnostics, Coralville, IA) the nodule was identified, and a ROI was selected including the nodule and a one-nodule-diameter radius of the surrounding lung (**Figure 4.1**). The bounding box coordinates of the ROI was recorded, so it could be replicated for the semi-automated assessment methods. Manual segmentation of these masks was performed using an in-house software developed in MatLab (Mathworks, Natick, MA)[11]. The manually generated masks included that of nodule (primary) and a secondary valid tissue mask. Valid tissue consisted of the nodule and surrounding parenchyma and excluded the chest wall, airways, and vessels interacting with the nodule or the parenchyma. The parenchyma masks (secondary) were calculated by subtraction of the nodule mask from the valid tissue mask.

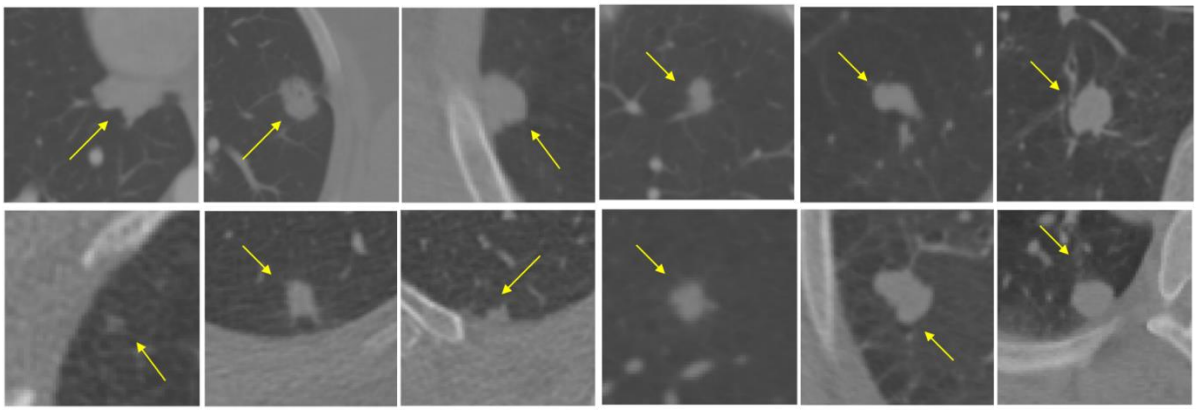

Figure 4.1: Sample ROI images with nodule location indicated by arrow.

4.2.3. Semi-automated Segmentation Tools

Five semi-automated segmentation methods were investigated and compared for accuracy, repeatability, and reproducibility. Functions from four image processing environments (FIJI-ImageJ (FIJI)[82], MeVisLab (MVL)[83], ITK-Snap (ITK-S)[84], Mukhopadhyay-MatLab (ML)[85]) were assembled into pipelines for semi-automated segmentations. One additional segmentation tool was developed in-house, Graph-cuts (GC). For each tool, a segmentation protocol was developed to ensure consistent use. Full descriptions of the pipelines are found in **Appendix C.1.** The segmentation method selected in this study was used in calculating the size-standardized parenchymal signal inclusion amount in the perinodular rings (inclusive) and band (exclusive). Further details on how these parenchyma masks were calculated is detailed in the **Appendix C.2.**

4.2.4. Analysis of Performance

Segmentation quality was assessed using well-established measures: sensitivity, specificity, Jaccard Distance (JD), Volumetric Error Rate (VER), and standardized Hausdroff Distance (SHD). The

results of these measures were linearly combined to create error and repeatability criteria. The equations for these measures and criteria are given in the **Appendix C.3**. The segmentation tools were evaluated using a weighted score of six criteria: the nodule segmentation and repeatability along with the user-grades for ease-of-use and ability to customize the tool. The Segmentation Error (40%) was calculated using a combination of the measures of similarity for each nodule. This measure was then ranked among the five tools such that each tool had 24 rankings, one for each nodule. The average of these rankings was taken as the criteria score. The Repeatability (20%) was calculated using the variances of the measures of similarity for each nodule in the same manner of Segmentation Error. Ease of use (15%) and Customization (25%) were user-indicated ratings based on interaction requirements, graphical user interface and intuitiveness, and post-segmentation processing abilities. A non-exclusive score of one to ten was given by each user, this score averaged and then multiplied by the weight factor (15% for Ease of use and 25% for Customization).

4.3. Results

4.3.1. Comparison of segmentation quality across tools

For each tool, two independent non-radiologist users of varying lung nodule segmentation experience implemented the segmentation pipeline five times for each of the nodules. Comparison of the three segmentation quality measurements (JD, VE, and SHD) was performed and summarized in **Figure 4.2**. The ML and GC were the highest performing tools, with similar values on the performance metrics; for ML and GC respectively the mean and standard deviation were; JD = 0.13±0.09, 0.15±0.07, VE = 0.06±.07, 0.05±0.09, SHD = 0.28±0.17, 0.28±0.19. FIJI achieved mid-range performance with regards to JD (0.24±0.17) and SHD (0.43±0.17) and was the only tool to consistently underestimate nodule volume (VE = -0.09±0.08). MVL also achieved sub-optimal performance in the three tested measures and had large variations in the JD and VE measures with lower variations in SHD shown by the standard deviations (JD = 0.09, VE = 0.21, and SHD = 0.08). Segmentations from ITK-S had the most edge-pixel difference of the segmentation tools coupled with a low nodule VE (0.1±0.17) of these masks, the errors tended to occur at edges that often visually represent vessels or pleural attachments. The trends across tools were consistent between nodule segmentations.

4.3.2. Results from criteria and scoring

The five tools were evaluated for segmentation error and repeatability as well as tool ease of use and customization using weighted scoring criteria (**Table 4.2**). The final score out of 100% indicates the tool's effectiveness and usability. The ease of use and customization are where FIJI, MVL, and ITK-S where sharply separated. FIJI received the lowest ease of use score of 6 as it required more than 20 user clicks per case and required 6 separate tabs of the graphical user interface to be open which affected the

ability of the user to quickly assess the nodule and parenchyma. Similarly, MVL and ITK-S performed poorly on the post-run segmentation modifications. MVL received the lowest score as it had no internal capability to customize the segmentation after a run, requiring the user to modify the segmentation outside of the tool. ITK-Snap did have internal capabilities to edit the segmentation post-run, however these modifications were time consuming and required user adjustments of energy cost function parameters. Using the weighted criteria as a guide we further analyzed the ML and GC approaches, testing for differences between error using the tool and inter-radiologist manual segmentation.

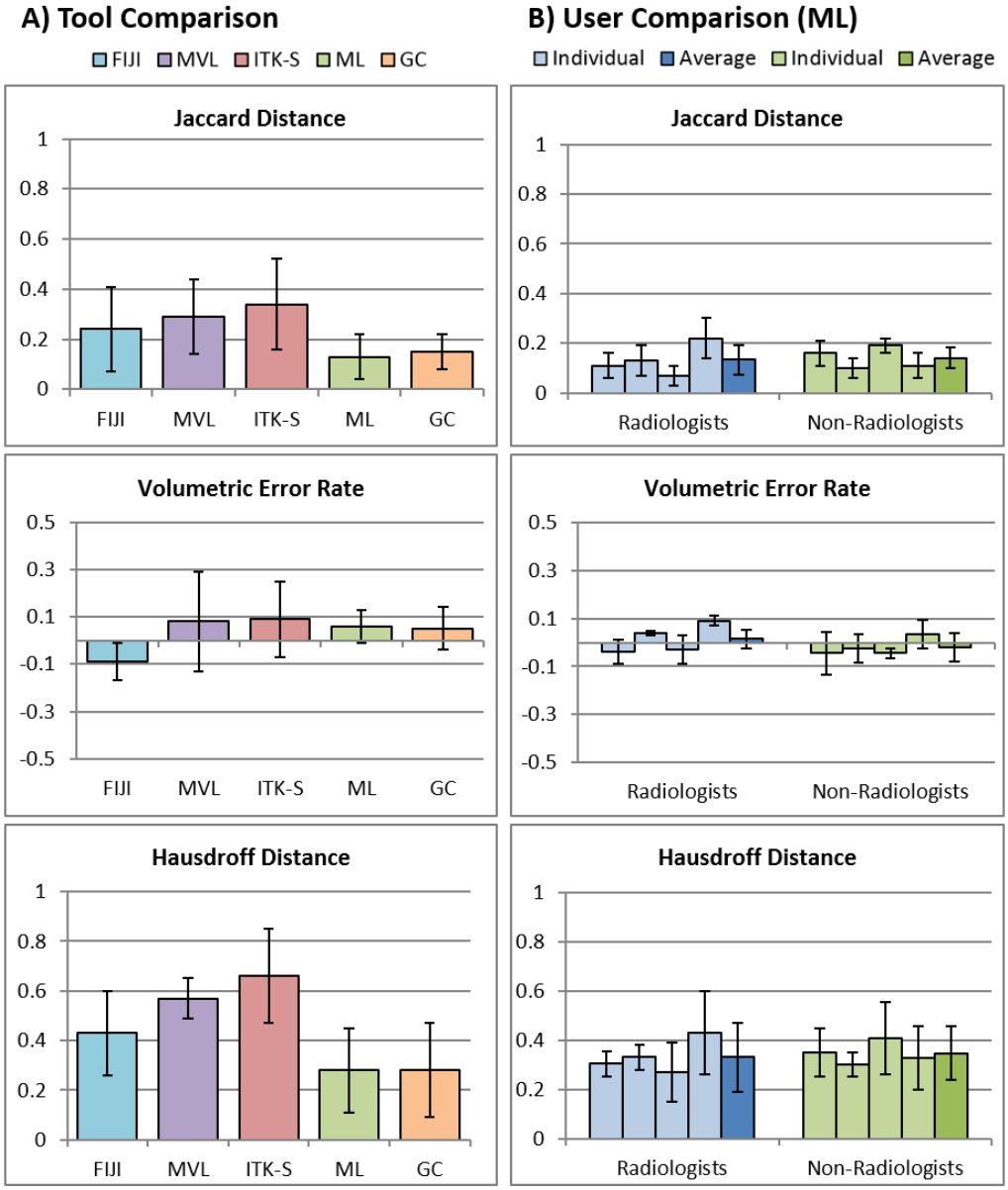

Figure 4.2: Segmentation results of three performance measures. A) Comparison between five semi-automated tools on full study cohort, B) Comparison of non-radiologists using ML to manual tracing of LIDC radiologists in the Variability Accuracy Cohort. Definition of abbreviations: FIJI – Fiji Is Just ImageJ; MVL – MeVisLab; ITK-S – ITK-Snap; ML - Mukhopadhyay-MatLab; GC – graph-cuts.

Table 4.2: Segmentation tool scoring and criteria. From cohort of 36 nodules, performance was calculated as the average of ten runs for each nodule.

Criteria	FIJI	MVL	ITK-S	ML	GC	Weight
Segmentation Error	8	7	6	9	9	40%
Repeatability	8	6	7	9	9	20%
Ease of Use	6	8	9	10	10	15%
Customization	10	3	5	10	9	25%
Weighted Score	**82%**	**60%**	**64%**	**94%**	**92%**	**100%**

Definition of abbreviations: FIJI – Fiji Is Just ImageJ; MVL – MeVisLab; ITK-S – ITK-Snap; ML - Mukhopadhyay-MatLab; GC – graph-cuts.

4.3.3. QIBA-Compliance Testing on Selected Tool

The guidelines produced by QIBA were used in order to provide a more standardized measure of the performance of the top scoring tool. As reported in Athelogou et al, 15% was the acceptable level of uncertainty in volumetric error outlined in the QIBA protocol.[46] The mean volumetric error for the segmentations of the 7 target tumors by ML tool was 3.54% **(Table 4.3).**

Table 4.3: Volumetric error rates for the 7 target tumors of the Lungman phantom using the ML segmentation method.

Shape	Nodule	Nominal Diameter	Volumetric Error
Spherical	**1**	10 mm	3.0%
	2	20 mm	4.2%
	3	40 mm	1.9%
Ovoid	**4**	10 mm	2.3%
	5	20 mm	2.2%
Lobulated	**6**	10 mm	6.2%
	7	20 mm	5.0%

4.4. Discussion

Studies using the LIDC's manual tracings from four expert radiologists have repeatedly shown variability between and within users, and this is a recognized limitation of manual approaches [44]. As non-negligible inter- and intra-reader variability can occur, there is a limitation in the resulting accuracy of any segmentation system based on the user's own subjective bias. Thereby, a tool achieving high accuracy compared to a single user's segmentation is potentially less powerful than a tool achieving high repeatability. By leveraging error with repeatability and including manual segmentations from non-radiologists and radiologists this study has attempted to lessen the directed bias that could result from focusing on a single user's segmentation as the truth standard. Having the semi-automated tools run by users of varying experience allowed us to test for overall-usability. The more experienced user had previously produced segmentations in all the segmentation tools, while the least experienced user had no segmentation exposure prior to the study.

It was the goal of this study to identify semi-automated approaches which could be used to create volumetric nodule segmentations in a more time efficient and less variable manner than manual editing,

aiming to strengthen the objective determination of longitudinal change. It is possible that to achieve the best accuracy of segmentation the nodule must first be categorized based on some criterion (shape, attachments, composition, etc.) and a different method of segmentation performed on each category. This prior knowledge of the individual nodule would be beneficial in selecting an algorithm that was optimized for that category; however, this was beyond the scope of this study. However, we did purposefully select nodules of varying characteristics to determine the tool's robustness to nodule identity; similarly Zhao et al and Athelogou et al compared the volumetric error and size agreement of segmentation tools to include breakdowns of location, shape, and edge characteristic effects to assess the effect on segmentation quality[46,77].

We developed a criterion scoring method **(Table 4.2)** to compare the tools based on qualities that were deemed important for our purposes: segmentation error, repeatability, ease of use, and customization. The weighting for each of these scores was based on relative importance to our segmentation goals; researchers with different priorities (i.e. not requiring customization) can adjust the relative weighting to determine a suitable tool. While the previous studies comparing semi-automated tools used measures of segmentation error and repeatability, they did not incorporate ease of use or customization assessment[46,77-79]. Customization is particularly important as a semi-automated tool cannot perform perfectly on all cases; therefore, the ability to alter the segmentation post-run is essential to utilizing a single tool for large scale nodule analysis. While the customization and ease of use measures proposed in this study are subjective to the reader, our measures incorporate both the assessed segmentation quality and the ability of the tool to re-assess a segmentation problem. Incorporating these measures into the scoring criteria provides a more comprehensive analysis of a tool's capability beyond accuracy and variability. As no segmentation algorithm is perfect, and with large-scale implementation some manual editing or correction is likely to be needed for difficult cases.

In this chapter, we systematically compared five semi-automated segmentation tools using a common cohort of chest CT scans with results shown on the tool's segmentation of pulmonary nodules. The best performing segmentation tool, ML, was used to segment the pulmonary nodules and surrounding parenchyma for machine learning pipelines discussed in **Chapter 5: QIC-RATE** and **Chapter 6: Application of QIC-RATE to Histoplasmosis Classification**.

CHAPTER 5: QIC-RATE

This chapter is adapted from the publication, "Optimized Perinodular Parenchyma Features Promote High Performance Machine Learning Tool for Lung Cancer", with accepted revisions in Medical Physics. Additional tool development exploratory studies are included in **Appendices D through G**.

5.1. Introduction

In **Chapter 3** we investigated the utility of previously developed post-imaging MPMs to stratify lung nodules based on cancer risk, concluding that while these models can be improved through local dataset calibration, room for risk assessment improvement persists. MPM risk assessment has not been widely adopted for clinical use, with the current clinical assessment of pulmonary nodules still based primarily on nodule size, composition, and growth information. Guidelines such as the American College of Radiology's Lung Imaging Reporting and Data System (Lung-RADS) criteria for pulmonary nodules identified with low-dose CT screening, the American College of Chest Physicians Evidence-Based Clinical Practice Guidelines, the British Thoracic Society Guidelines and the Fleischner society guidelines for incidental nodules use the size of the lung nodule as a key indicator to determine appropriate clinical follow-up procedures[31,33,51,52].

By only using clinician-collected variables, there is the potential that not all diagnosis-informative data collected is being used to the full potential. CT scans contain a wealth of potential information with standardized base units of size (mm) and image value (HU). This standardized acquisition improves the quality and reproducibility of automatic quantitative imaging characteristics (QICs) that can be calculated from the image data. Risk assessment tools, developed through machine learning algorithms, can utilize QICs extracted directly from the CT scans, such as nodule shape and texture, to differentiate between malignant and benign disease states[10-25]. The traditional focus of imaging-based risk models for lung cancer has been on nodule and border features with a range in area-under the receiver operator curve (AUC-ROC) performance (0.821 - 0.99)[11-16].

The perinodular parenchyma has biological importance with respect to cancerous changes such as cell migration, inflammation, and vascularization. Morphological characteristics from this region including spiculation and structural distortion of the parenchyma have been reported as indicative of malignancy, improving observer performance, and included in the MPMs[57,59,60,86]. Recently, improvements in lung nodule classification have been demonstrated through the inclusion of perinodular parenchymal QICs using traditional machine learning (AUC-ROC: 0.938[10] and 0.915[24]) and deep learning methods have indirectly examined parenchymal inclusion (AUC-ROC of 0.899 to 0.946)[20-22], but the degree to which parenchyma has been included has varied. From the reported literature, it is not clear

where the optimal parenchymal radiological signal resides, and how these features interplay with features from the nodule and its borders.

In this chapter, we develop a pipeline for machine learning tool development which utilizes features from the nodule and size-standardized regions of the surrounding parenchyma, termed Quantitative Imaging Characteristics for Risk Assessment from the Tumor and Environment (QIC-RATE). We systematically investigate the optimal regions of perinodular parenchyma to use for feature extraction and classification. With a focus on feature transparency, we explore the trends within and between regions of perinodular parenchyma by nodule- standardized parenchymal quartile-bands. Finally, we compare the value QIC-RATE could have on follow-up pathways versus the established Fleischner Society guidelines in an independent validation dataset.

5.2. Materials and Methods

A systematic processing pipeline was developed to identify the optimal feature set for QIC-RATE, and independent validation testing, as depicted in **Figure 5.1**. The QIC-RATE tool development involved the selection of a feature set and classifier training, while validation was executed on the top performing candidate tool. The QIC-RATE tool development pipeline was built on a foundation of prior work from our lab[10,11]. The prior approach provided proof of the added benefit to computer-aided diagnosis of including perinodular signal over nodular signal alone. As more subjects provide more data about the true variability in the population, it is necessary to improve upon the mechanisms utilized in the prior work for use on a larger cohort. **Table 5.1** summarizes the prior approach, the potential challenges identified in the methodology when adding additional cases, the solutions tested, and the final selected mechanism for QIC-RATE.

5.2.1. Study Cohorts

Subjects included 363 pulmonary nodules, ≤30mm in diameter (74 malignant, 289 benign) from three study data sources: COPDGene, NLST, and the SPIE LungX Challenge [26,40,47]. The prior approach used a subset of 50 subjects from the current 363 cohort to demonstrate the value of feature extraction from the lung parenchyma[10,11]. An independent validation cohort of 100 pulmonary nodules (50 malignant, 50 benign) from the INHALE study was used to test QIC-RATE[41]. Further demographics and scanning parameters for the two cohorts is described in **Table 5.2**. For more complete details on the origin datasets, please see **Chapter 2.**

Table 5.1: Identified areas of improvement in prior approach, solutions tested as part of this dissertation, and the selected approach that was implemented in the final QIC-RATE system.

	Prior Approach[11]	Challenges Identified	Solutions Tested	Selected Approach
Segmentation (Chapter 4)	Manual nodule and parenchyma segmentation	• Time intensive • Prone to inter- and intra- user variability	Semi-automated segmentation methods 1-5	ML method
Feature Extraction (Appendix D)	304 nodule and parenchyma features describing image intensity and texture, and nodule border, shape, and size	Improvement through inclusion of other features deemed potentially predictive in other literature	Additional features: • Intensity • GLRL • GLSZ • NGTD • Size/Shape	Extraction of all features in **Tables 5.3-4**
Set Reduction (Appendix E)	Statistical significance	Highly correlated features persist which can affect model stability	• K-medoids • PCA	K-medoids using k=adjusted best silhouette
Set Selection (Appendix F)	Leave-one-out feed-forward feature selection from ANN	• Time intensive • Selection linked to classifier (increases overtraining probability)	• K-medoids • Majority Votes • Mutual Information Measures • Random Forest Importance	Information optimization (IO)
Classification (Appendix G)	ANN • 2 hidden layers • Hyperparameters • Up to N/10 features	Improvement by: • different classifier • hyperparameter setup • Number of features allowed for selection	• Support Vector machine • Conditional inference trees • Ensembling • Jittering hyperparameters • Up to N/5 features	• Ensemble Artificial Neural Network • Built with jittered hyperparameters • On up to N/5 features

Definition of Abbreviations: ML – Mukhopadhyay-MatLab; GLRL – gray-level run-length; GLSZ – gray-level size-zone; NGTD – neighborhood gray-tone difference; ANN – artificial neural network; PCA – principle component analysis.

The development cohort (N=363) was diverse in subject demographics, scanner parameters, and CT manufacturer; statistical difference in subject demographics existed between malignant and benign nodules (**Table 5.2**). As this cohort was established by combination of different parent academic studies, we explored the number of subjects that would have met lung cancer low-dose CT (LDCT) screening eligibility criteria based on age and smoking pack-years. Scanning parameters in the development cohort were in accordance with recommended protocols for high-resolution CT studies [87], with the exception of the NLST cases which was a LDCT protocol (reconstructed thin slice thickness). The demographics and scanning parameters for the validation cohort (N=100) were more evenly matched between malignant and benign cases due to the nature of the INHALE study's inclusion criteria (**Table 5.2**).

Table 5.2: Demographic and scanning parameters for the QIC-RATE lung nodule study.

		Malignant	Benign	p
Development/Testing	Subjects	74	289	-
	Study (COPDGene:NLST:LungX)	30:6:38	239:8:42	< 0.01
	Age, yrs (mean±SD)	65.5 ± 11.3	53.2 ± 13.0	< 0.01
	Sex (Female: Male)	34:40	179:110	< 0.01
	Pack-years*, yrs (mean±SD)	37.7 ± 30.4	10.7 ± 15.7	< 0.01
	Nodule size, mm Range (mean±SD)	5-30 (13.6 ± 6.2)	4-30 (7.79 ± 13.3)	< 0.01
	Nodule size ≤ 15mm	50	240	< 0.01
	LDCT screening eligible* (Yes: No)	33:3	69:178	< 0.01
	Kilovoltage (range, mean)	120-120 kVp, 120 kVp	120-120 kVp, 120 kVp	1.00
	Tube Current (range, mean)	60-440 mA,† 339 mA	40-500 mA,† 330 mA	0.89
	Slice thickness (range, mean)	0.6-1.3 mm, 0.8 mm	0.6-1.3mm, 0.7mm	0.97
	CT Manufacturer (GE:Philips:Siemens)	19:35:20	86:97:106	<0.01
Validation	Subjects	50	50	-
	Study (INHALE)	50	50	1.00
	Age (mean±SD)	64.0 ± 10.7	62.5 ± 10.9	0.46
	Sex (Female: Male)	35:15	31:19	0.34
	Pack-years*, years (mean±SD)	33.5 ± 30.1	30.1 ± 23.6	0.51
	Nodule size, mm Range (mean±SD)	5-30 (19.9 ± 7.4)	9-30 (13.66 ± 4.8)	< 0.01
	Nodule size ≤ 15mm	17	35	< 0.01
	LDCT screening eligible (Yes: No)	18:32	16:34	0.67
	Kilovoltage (range, mean)	120-120 kVp, 120 kVp	120-120 kVp, 120 kVp	1.00
	Tube Current (range, mean)	160 – 351 mA, 237 mA	160 – 386 mA, 265 mA	0.62
	Slice thickness (range, mean)	0.6-0.8 mm, 0.7 mm	0.6-0.8 mm, 0.7 mm	0.98
	CT Manufacturer (GE:Philips:Siemens)	17:19:14	16:22:12	0.82

Definition of abbreviations: SD - standard deviation; LDCT - low dose computed tomography; GE - General Electric; *: smoking pack-year data was not available for the LungX Challenge; †: low-dose NLST scans included were reconstructed to higher resolution at time of acquisition.

5.2.2. Segmentation of Nodule and Parenchyma

The nodule and parenchyma were semi-automatically segmented into ROI using a modified version of the proposed pipeline by Dhara et al. [85,88]. For more complete details on the segmentation tool select process, please see **Chapter 4: Segmentation**. The nodule mask was grown using a binary image dilation to produce parenchyma quartile-bands: 25%, 50%, 75%, and 100% of the maximum in-plane diameter of the nodule (**Figure 5.1**). Candidate QIC-RATE tools were created using a nesting pattern of the bands: Nodule (no bands), Margin (Nodule + 25% band), Immediate (Nodule + 25% band + 50%

band), Extended (Nodule + 25% band + 50% band + 75% band), and Extended+ (Nodule + 25% band + 50% band + 75% band + 100% band).

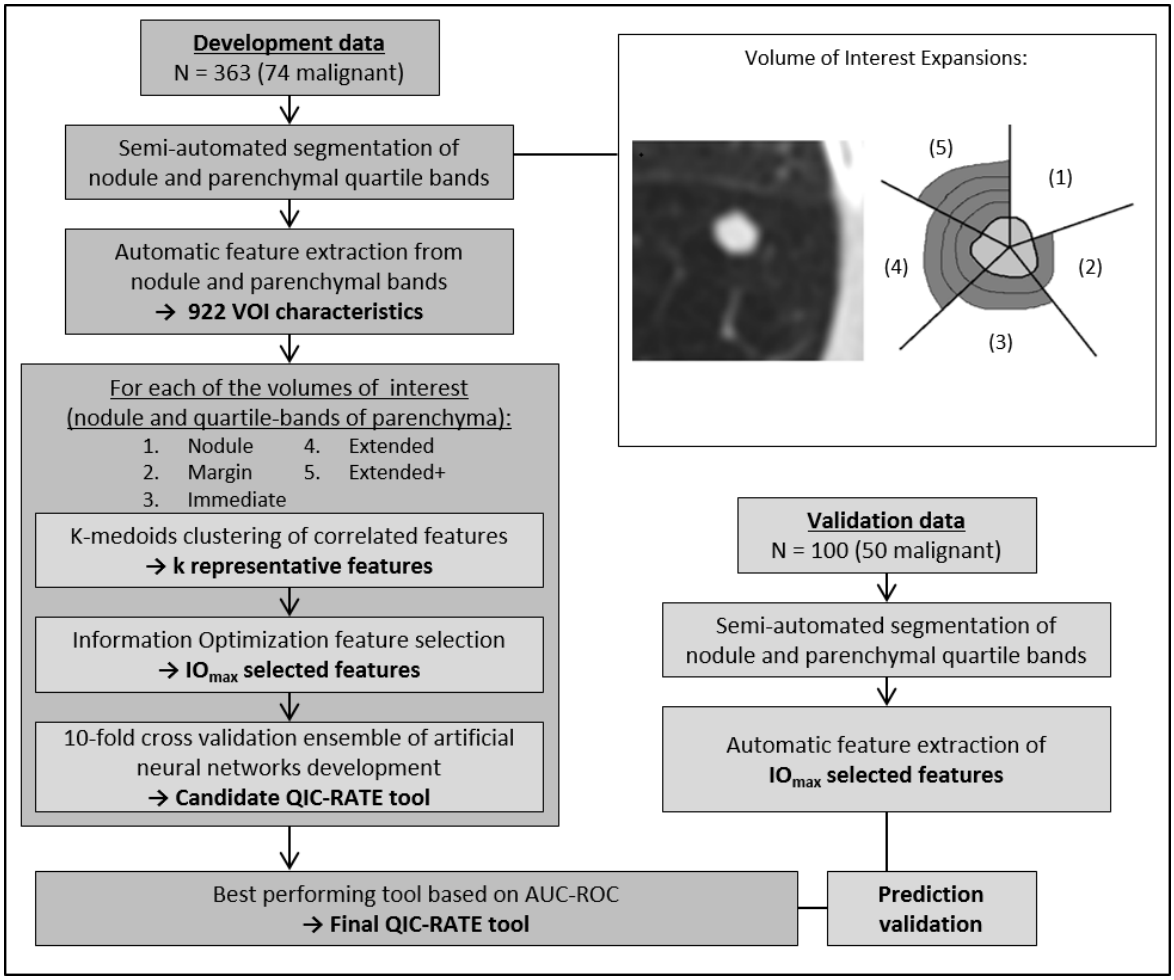

Figure 5.1: Overview of QIC-RATE tool development and validation pipeline. The depiction of the varying amounts of parenchyma tested through the pipeline (in quartile-bands) include; (1) Nodule, (2) Margin [nodule, 25%], (3) Immediate [nodule, 25%, 50%], (4) Extended [nodule, 25%, 50%, 75%], (5) Extended+ [nodule, 25%, 50%, 75%, 100%]. Definition of abbreviations: QIC-RATE - , volume of interest (VOI), information objective function maximum point (IOmax), area under the receiver operating characteristic curve (AUC-ROC).

5.2.3. Development of QIC-RATE

Quantitative CT features were automatically extracted from the nodule and parenchyma quartile volumes to produce candidate QIC-RATE tools (**Figure 5.1**, **Table 5.3-5.4**). These included 14 volumetric measures of intensity histogram (IH), 136 volumetric Law's energy measures (LTEM), 13 volumetric gray-level run-length measures (GLRL), 13 volumetric gray-level size-zone measures (GLSZ), 5 volumetric neighborhood gray-tone difference measures (NGTD), and 17 measures of size and volumetric shape (SzSp) including 11 border measures (BASC and BCRP) [10,89-94]. For full overview on these QICs, see **Appendix D**.

Features were clustered based on pair-wise correlations using the k-medoids method resulting in k-clusters with k-representative medoid features[95,96]. Determination of k was done by optimization of the average cluster silhouette with the method adjusted to not penalize for clusters of one feature. For a full summary of the tests used to determine this method of feature reduction, see **Appendix E**. From the reduced group of medoids, a set of predicting features was selected using an objective function of information theory measures[97]. The maximum selected set size was determined from the information objective function maximum point (IO_{max}). For a full summary of the tests used to determine this method of feature set selection, see **Appendix F**. In cases where the IO_{max} was larger than one predictor for every five to ten cases the set size was capped at 72 features[98]. The selected feature sets were used to train the ANNs with performance analyzed through 10-fold kCV (k-fold cross validation, see **Appendix A.3**). As random initialization of weights in ANN development can affect classifier performance, we further developed an ensemble of ten ANNs (ENN) for final prediction values. For full details of classifier method selection and training, see **Appendix G**.

5.2.4. Performance and Comparison

Detailed information on the specific performance measures is included in the **Appendix A.** In brief, tool performance was assessed using AUC-ROC (Delong) and AUC-PR. To determine the statistical advantage of one candidate tool over another on a given dataset, we employed Delong and McNemar's (Youden J statistic threshold) tests. The potential impact on clinical follow-up response was assessed by comparing the predictions from the QIC-RATE tool to the Fleischner Society Pulmonary Nodule Follow-up Guidelines.

Table 5.3: Size and shape features extracted from the nodule ROI.

Feature Groups	Features
Border Centroid Radial Rays (BCRR)	• Mean of Border • Standard Deviation of Border • Mean of Slopes • Standard Deviation of Slopes • Mean of Columns • Standard Deviation of Columns
Border Absolute Sphere Comparison (BASC)	• Mean • Variance • Kurtosis • Skewness • Range
Whole Tumor Characteristics (WTC)	• Sphericity • Maximum in-plane diameter (RECIST) • Radius • Volume * • Equivalent H_20 Area *,% • Equivalent H_20 Diameter *, %

***: new feature; %: CT-specific feature**

Table 5.4: Intensity and texture features extracted from the nodule and the perinodular parenchyma ROIs.

Feature Group	Features
Intensity Histogram (IH)	• Mean • Variance • Maximum • Minimum • Median • Full-Width-at-Half-Maximum • Entropy • Kurtosis • Skewness • 5^{th} Percentile * • 95^{th} Percentile * • 25^{th} Percentile * • 75^{th} Percentile * • Proportion over 100 HU *, %
Law's Texture (LTEM)	• Mean Energy Measures (14 2D, 34 3D) • Variance Energy Measures (14 2D, 34 3D) • Kurtosis Energy Measures (14 2D, 34 3D) • Skewness Energy Measures (14 2D, 34 3D)
Gray Level Size Zone Texture (GLSZ)	• Small Zone Emphasis * • Large Zone Emphasis * • Gray-Level Non-uniformity * • Zone-Size Non-uniformity * • Zone Percentage * • Low Gray-Level Zone Emphasis * • High Gray-Level Zone Emphasis * • Small Zone Low Gray-Level Emphasis * • Small Zone High Gray-Level Emphasis * • Large Zone Low Gray-Level Emphasis * • Large Zone High Gray-Level Emphasis * • Gray-Level Variance * • Zone-Size Variance *
Neighborhood Gray Tone Difference (NGTD)	• Coarseness * • Contrast * • Busyness * • Complexity * • Strength *
Run Gray Level Length Texture (GLRL)	• Short Run Emphasis * • Long Run Emphasis * • Gray-Level Non-uniformity * • Run-Length Non-uniformity * • Run Length Percentage * • Low Gray-Level Run Emphasis * • High Gray-Level Run Emphasis * • Short Run Low Gray-Level Emphasis * • Short Run High Gray-Level Emphasis * • Long Run Low Gray-Level Emphasis * • Long Run High Gray-Level Emphasis * • Gray-Level Variance * • Run-Length Variance *

***: new feature; %: CT-specific feature**

5.3. Results

5.3.1. QIC-RATE Performance

The best performing candidate tool included the nodule and surrounding tissue from the 25%, 50%, and 75% quartile-bands (Extended QIC-RATE). In the development cohort, the Extended QIC-RATE tool achieved an AUC-ROC of 1.0 – or complete separation of the classes along the ensemble-ANN decision boundary and achieved the highest AUC-PR (0.945). The performance of the undivided, inclusive (border-to-75%) ring was also calculated and achieved weaker measures (AUC-ROC=0.938, AUC-PR=0.916). On the independent validation cohort, the Extended tool achieved an AUC-ROC=0.965 (accuracy 98%, sensitivity 100%, specificity 96%). Delong and McNemar's test comparisons showed the four tools incorporating parenchymal signal (Margin, Immediate, Extended, Extended+) were statistically better than the Nodule tool ($p < 0.01$); there was no statistical difference between the candidate tools developed with parenchymal features discriminatory power. **Figure 5.2** shows a visual representation of the feature set selection over the five tool candidates, demonstrating the number of perinodular parenchyma features included in the candidate tools and the wealth of feature types selected.

For the Extended tool the objective function's IO_{max} of 76 predictors was adjusted to 72 predictors to prevent overfitting, to produce the Extended and Extended+ tools (**Table 5.5**). To test if fewer selected features were needed to maintain this boundary plane we built ensemble-ANNs for each feature set size between 2 and 72 predictors. Complete class separation was achieved for all ensemble-ANNs between 50 and 72 predictors (AUC-ROC = 1.0, AUC-PR = 0.945). The remainder of the results will focus on the Extended tool.

Table 5.5: Performance results from 10-fold cross validation on the development cohort of QIC-RATE candidate tools.

Candidate	Set Size	AUC-ROC	AUC-PR	Youden	Sensitivity	Specificity
Nodule	22	0.919	0.891	0.36	0.85	0.92
Margin	38	0.982	0.916	0.28	0.90	0.98
Immediate	55	0.998	0.943	0.40	0.93	0.98
Extended	**50**	**1.000**	**0.945**	0.38	**1.00**	**1.00**
Extended+	72	0.998	0.944	0.35	0.93	0.98

Definition of abbreviations: AUC-ROC – area-under-curve of receiver-operator characteristic; AUC-PR – area-under-curve of precision-recall

5.3.2. Extended QIC-RATE Feature Set

The fifty selected features included 13 IH, 15 GLRL, 12 GLSZ, 5 NGTM, 1 ASC, and 3 SzSp. The region from which the included features were extracted included 19 nodular, 2 of band 25%, 12 of band 50%, and 17 of band 75%, such that the final model contained more features extracted from the parenchyma than from the nodule. In the development cohort, cutoff values of 0.38 from the receiver operating characteristic prioritize the correct classification of malignant cases (**Figure 5.3**). Application of the

Extended QIC-RATE in the independent validation cohort resulted in an accuracy of 98%, with no misclassification of malignant cases.

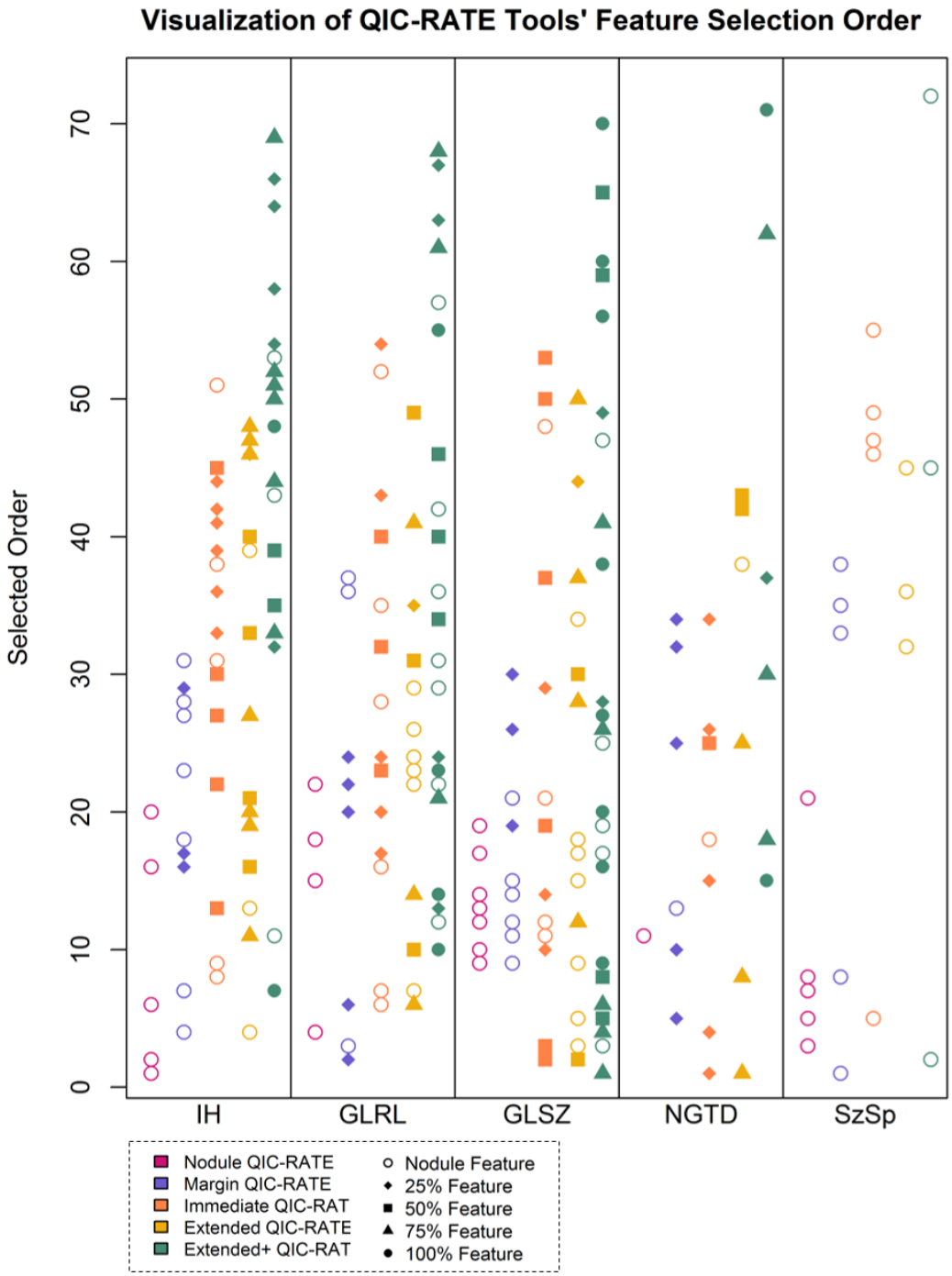

Figure 5.2: Visualization of feature selection by category for each of the five candidate tools – object color indicates the candidate tool and object shape indicates the ROI of feature extraction. Definition of abbreviations: IH – intensity histogram; GLRL – gray-level run-length; GLSZ – gray-level size-zone; NGTD – neighborhood gray-tone difference; SzSp – size and shape

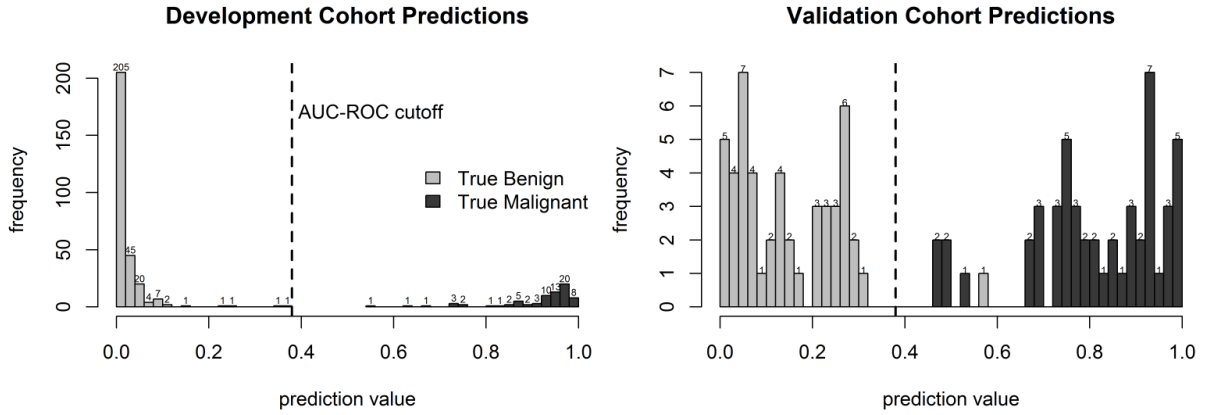

Figure 5.3: The Extended QIC-RATE resulted in complete division of malignant and benign lung nodule cases in cross validation testing of the development cohort (363 cases), with a threshold of 0.38 as determined using the Youden's J statistic. Output lung cancer risk prediction values range from 0 (confidently benign) to 1(confidently malignant).

A complete list of the fifty selected features for the Extended QIC-RATE tool with mean, standard deviation, p-value (from t-test or Wilcoxon rank sum test as appropriate), and Pearson's correlation with nodule size is included (see **Table 5.6**) which lists the features selected with statistics. To summarize, the first five features selected included two features from the surrounding parenchyma quartile-bands, followed by three nodule features. Selected first was a NGTD feature describing the coarseness of texture in the 75% parenchyma quartile-band, which is a high order feature where large values represent areas where the gray-tone differences are small, therefore leading to a high degree of local uniformity in intensity. Malignant cases had lower values (0.005±0.013) than benign cases (0.011±0.017) and this was statistically different (p = 0.023). Next, feature selection chose a GLSZ texture feature indicating large zone emphasis in the 50% parenchyma quartile-band. This feature multiplies each zone by the size of the zone squared, thus high values imply large zones within the texture. Again, malignant cases had lower values (159.5±386.4), than benign (3087.1±8332.7), this was statistically significant (p=0.007). The third selected feature was GLSZ extracted from the nodule indicating small zone low gray-level emphasis with malignant cases having lower values (0.026±00.23) than benign (0.036±0.025) at a significant level (p<0.001). This feature is larger when there is an emphasis of small zones of low intensity within the texture. The next two selected features were also nodule-based being the entropy (IH) and high gray-level zone emphasis (GLSZ).

Table 5.6: Fifty features selected for use in Extended QIC-RATE tool, including p-value and correlation with nodule size.

#	Feature	Malignant		Benign		p	r-size
		Mean	SD	Mean	SD		
1	75% Coarse Texture	5.40E-03	1.36E-02	1.04E-02	1.65E-02	0.02	-0.42
2	50% Large Zone Emphasis	1.60E+02	3.86E+02	3.09E+03	8.33E+03	< 0.01	0.37
3	Nodule Small Zone Low GL Emphasis	2.61E-02	2.30E-02	3.61E-02	2.45E-02	< 0.01	-0.30

Table 5.6, continued: Fifty features selected for use in Extended QIC-RATE tool.

#	Feature	Malignant		Benign		p	r-size
		Mean	SD	Mean	SD		
4	Nodule Entropy HU	7.38E+00	9.30E-01	7.25E+00	1.03E+00	0.33	0.32
5	Nodule High GL Zone Emphasis	2.55E+02	1.16E+02	2.52E+02	7.54E+01	0.34	-0.16
6	75% Long Run High GL Emphasis	1.79E+02	9.12E+01	1.40E+02	7.81E+01	< 0.01	-0.05
7	Nodule Run Length Percentage	9.19E-01	5.98E-02	9.02E-01	9.72E-02	0.43	-0.13
8	75% Contrast Texture	2.55E-01	2.49E-01	1.25E-01	1.38E-01	< 0.01	-0.11
9	Nodule GL Variance Zones	5.61E-02	2.03E-02	4.79E-02	1.45E-02	< 0.01	0.18
10	50% Run length Variance	2.40E-04	1.37E-04	2.80E-04	9.92E-05	< 0.01	-0.29
11	75% 95^{th} Percentile HU	-9.86E+02	4.50E+01	-9.60E+02	4.36E+01	< 0.01	-0.17
12	75% Zone Size Variance	6.99E-05	1.18E-04	1.25E-04	1.20E-04	< 0.01	-0.37
13	Nodule Kurtosis HU	6.65E+00	9.08E+00	5.47E+00	7.24E+00	< 0.01	0.11
14	75% GL Run Variance	2.34E-02	2.56E-02	2.06E-02	3.47E-02	0.02	-0.20
15	Nodule Large Zone Low GL Emphasis	1.36E+01	4.79E+01	1.41E+02	4.89E+02	0.52	0.00
16	50% Variance HU	3.36E+04	1.61E+04	2.20E+04	1.25E+04	< 0.01	0.00
17	Nodule Run Length Variance	1.78E-04	2.07E-04	2.45E-04	3.42E-04	0.04	-0.19
18	Nodule Large Zone High GL Emphasis	6.91E+04	3.37E+05	5.43E+04	2.04E+05	0.46	0.04
19	75% Full-Width-at-Half-Maximum	4.90E-02	5.41E-02	6.34E-02	4.16E-02	< 0.01	0.29
20	75% Mean HU	-7.79E+02	8.79E+01	-8.12E+02	7.87E+01	< 0.01	-0.09
21	50% Maximum HU	-2.20E+02	5.63E+01	-2.37E+02	6.89E+01	< 0.01	0.30
22	Nodule Short Run Low GL Emphasis	2.51E-02	2.19E-02	3.85E-02	2.60E-02	< 0.01	-0.26
23	Nodule Long Run Low GL Emphasis	3.46E-02	3.98E-02	7.45E-02	1.05E-01	< 0.01	-0.06
24	Nodule GL Variance Runs	8.49E-02	8.18E-02	7.09E-02	5.53E-02	0.35	-0.21
25	75% Busyness Texture	2.75E+00	3.21E+00	2.55E+00	4.00E+00	< 0.01	0.64
26	Nodule High GL Run Emphasis	2.49E+02	1.40E+02	2.40E+02	1.17E+02	0.87	-0.17
27	75% Median HU	-8.17E+02	9.00E+01	-8.45E+02	7.71E+01	< 0.01	-0.11
28	75% GL Non-uniformity Zones	4.68E-02	1.18E-02	5.28E-02	1.27E-02	< 0.01	-0.16
29	Nodule GL Non-uniformity Runs	7.67E-02	4.12E-02	7.26E-02	4.65E-02	0.09	0.13
30	50% Small Zone High GL Emphasis	2.21E+02	8.03E+01	1.89E+02	6.84E+01	< 0.01	-0.13
31	50% GL Variance Run	3.93E-02	6.29E-02	3.15E-02	3.76E-02	0.01	-0.28
32	Equivalent H_20 Area Centroid Slice	1.25E+03	1.05E+03	1.11E+03	1.28E+03	0.02	0.45
33	50% Entropy HU	8.58E+00	7.17E-01	8.03E+00	7.43E-01	< 0.01	0.30
34	Nodule Large Zone Emphasis	4.44E+02	1.09E+03	1.71E+03	4.86E+03	0.36	0.02
35	25% GL Non-uniformity Run	4.70E-02	3.74E-02	5.65E-02	3.52E-02	< 0.01	0.02
36	Nodule Sphericity	5.32E-01	1.62E-01	5.94E-01	1.94E-01	< 0.01	-0.48
37	75% GL Variance Zones	4.26E-03	1.85E-02	3.64E-03	1.13E-02	0.02	-0.24
38	Nodule Contrast Texture	4.06E-01	8.68E-01	5.03E-01	6.10E-01	0.02	-0.40
39	RECIST Diameter	1.36E+01	6.20E+00	7.87E+00	1.33E+01	< 0.01	1.00
40	50% Full-Width-at-Half-Maximum	4.46E-02	5.13E-02	5.72E-02	3.90E-02	< 0.01	0.37
41	75% Low GL Run Emphasis	6.17E-02	4.66E-02	4.97E-02	3.83E-02	0.06	< 0.01
42	50% Contrast Texture	3.68E-01	2.97E-01	2.15E-01	1.95E-01	< 0.01	-0.25
43	50% Complexity Texture	1.19E+03	4.65E+02	8.86E+02	3.66E+02	< 0.01	-0.16
44	25% Long Run Emphasis	1.56E+01	6.37E+01	5.86E+02	1.80E+03	0.21	0.13
45	Mean Absolute Sphere Comparison	4.14E-01	2.87E-01	4.54E-01	3.30E-01	0.73	0.36
46	75% 25^{th} Percentile HU	-9.96E+02	3.88E+01	-9.73E+02	4.15E+01	< 0.01	-0.20
47	75% Maximum HU	-2.35E+02	9.71E+01	-2.66E+02	1.21E+02	< 0.01	0.28
48	75% Skewness HU	1.49E+00	8.84E-01	2.08E+00	9.62E-01	< 0.01	0.17
49	50% GL Non-Uniformity Run	5.63E-02	1.97E-02	7.88E-02	3.67E-02	< 0.01	0.28
50	75% Small Zone Low GL Emphasis	2.14E-02	1.23E-02	1.89E-02	1.12E-02	0.13	-0.04

Definition of abbreviations: SD – standard deviation; HU – Hounsfield unit; GL – gray level; r-size – Pearson's correlation with nodule RECIST diameter

Presented in **Table 5.7** are the features used in the Extended QIC-RATE tool, which were selected from more than one location (i.e. nodule and parenchyma, or different quartile-bands of parenchyma), which was 23/50 features. While these features are extracted in the same manner, the spatial location of the extracted region is effective. Large values in entropy features indicate a large amount of randomness in gray levels of the ROI. Full-width-at-half-maximum (FWHM) of the histograms of parenchyma bands tended to be smaller in malignant cases indicating a thinner, more peaked histogram shape. The gray-level non-uniformity demonstrates while the malignant nodule showed increased non-uniformity, the tissue surrounding the malignant nodules tended to be lower than their benign counter-parts. Run length variance tended to be lower in malignant cases indicating more homogeneous runs. The small zone low gray-level emphasis demonstrated a converse effect with malignant nodules tending to have lower values while the tissue surrounding those nodules obtained a mean higher than that of benign nodule's surrounding tissues. Similarly, the contrast in texture showed a converse effect between nodule and parenchyma signal. The high-order feature is high in the surrounding malignant nodules indicating increased amount of local variation in intensity. On the other hand, contrast texture tends to be lower in the nodule proper indicating a smaller amount in local intensity variation. While there was a size bias in the development cohort (malignant: 13.6mm±6.2, benign: 7.8mm±13.3, $p<0.001$), the maximum in-plane diameter was selected later (39/50) in the Extended QIC-RATE. In addition, on nodules with size $\leq$ 15mm, the Extended QIC-RATE tool maintained high performance in both development (AUC-ROC=1.0, AUC-PR=0.943) and validation (AUC-ROC=0.998, AUC-PR=0.877).

5.3.3. Fleischner Society Guidelines Comparison

We analyzed the potential effect the QIC-RATE tool would have on the follow-up response compared to the Fleischner Society Pulmonary Nodule Follow-up Guidelines as the INHALE study used for validation was not a lung cancer screening cohort (see **Table 5.8**, which provides the results of applying the guidelines to the validation)[41]. These guidelines are stratified by size and nodule composition, as all nodules in this study were solid, we can separate into three categories: Category-1: CT in 12 months, Category-2: CT in 6-12 months, and Category-3: CT, biopsy, or positron emission tomography in 3 months. No size distribution criterion was enforced on nodule inclusion in this study. As such, 97% of the validation cohort fell into the third size-based category; this differed from the development cohort where the split was more balanced (Category 1: 23%, Category 2: 35%, Category 3: 41%). The Extended QIC-RATE tool identified 50 malignancies that would have required follow-up with a waiting period of 3-12 months[51]. The Extended QIC-RATE tool also recognized 48 benign nodules that would have required a 3-month follow-up with imaging or biopsy. We demonstrate the potential acceleration of malignant follow-ups over the Fleischner guidelines; for three malignant cases, the

Fleischner Society guidelines would have recommended a CT in 6-12 months while the Extended QIC-RATE tool would immediately send these patients to treatment. Similarly, for an additional 45 subjects with malignant nodules, the guidelines would have recommended a follow-up in 3 months of imaging and/or biopsy.

Table 5.6: Example Feature trends in malignant nodules from Extended QIC-RATE.

Group	Feature	#	Trend	Malignant		Benign		p
				Mean	SD	Mean	SD	
IH	Nodule Entropy	4	↑	7.38	0.93	7.25	1.03	0.33
	50% Entropy	33	↑	8.58	0.72	8.03	0.74	<0.01
	50% Maximum HU	21	↑	-220.4	56.3	-236.6	68.9	<0.01
	75% Maximum HU	47	↑	-235.0	97.1	-265.6	120.7	<0.01
	50% Full-Width-at-Half-Maximum	40	↓	0.04	0.05	0.06	0.04	<0.01
	75% Full-Width-at-Half-Maximum	19	↓	0.05	0.05	0.06	0.04	<0.01
GLRL	Nodule GL Non-uniformity Runs	29	↑	0.08	0.04	0.07	0.05	0.09
	25% GL Non-uniformity Run	35	↓	0.05	0.04	0.06	0.04	<0.01
	50% GL Non-uniformity Run	49	↓	0.06	0.02	0.08	0.04	<0.01
	Nodule Run Length Variance	17	↓	1.8E-04	2.1E-04	2.5E-04	3.4E-04	0.04
	50% Run length Variance	10	↓	2.4E-04	1.4E-04	2.8E-04	9.9E-05	<0.01
	Nodule GL Variance Runs	24	↑	0.08	0.08	0.07	0.06	0.35
	50% GL Variance Run	31	↑	0.04	0.06	0.03	0.04	0.01
	75% GL Variance Run	14	↑	0.02	0.03	0.02	0.03	0.02
GLSZ	Nodule GL Variance Zones	9	↑	0.06	0.02	0.05	0.01	<0.01
	75% GL Variance Zones	37	↑	4.3E-03	1.9E-02	3.6E-03	1.1E-02	0.02
	Nodule Large Zone Emphasis	34	↓	443.6	1089.9	1708.6	4862.4	0.36
	50% Large Zone Emphasis	2	↓	159.5	386.4	3087.1	8332.7	0.01
	Nodule Small Zone Low GL Emphasis	3	↓	0.03	0.02	0.04	0.02	<0.01
	75% Small Zone Low GL Emphasis	50	↑	0.02	0.01	0.02	0.01	0.13
NGTD	Nodule Contrast Texture	38	↓	0.41	0.87	0.50	0.61	0.02
	50% Contrast Texture	42	↑	0.37	0.30	0.22	0.19	<0.01
	75% Contrast Texture	8	↑	0.26	0.25	0.12	0.14	<0.01

Definition of abbreviations: SD – standard deviation; HU – Hounsfield unit; GL – gray level; IH – intensity histogram; GLRL – gray level run length; GLSZ – gray level size zone; NGTD – neighborhood gray-tone difference

Table 5.7: Extended QIC-RATE tool compared to Fleischner Society Pulmonary Nodule Follow-up Guidelines.

Fleischner Size-based Recommendations		Malignant		Benign		Time reduction (months)
		Pathology	QIC-RATE-prediction	Pathology	QIC-RATE-prediction	
< 6 mm	FFU: CT in 12 months	1	1	0	0	12
6 to 8 mm	FFU: CT in 6-12 months	2	2	0	0	12-24
> 8 mm	FFU: CT, Biopsy, or PET-CT in 3 months	47	47	50	48	285

Definition of abbreviations: FFU – Fleischner Society follow-up recommendations for solitary pulmonary nodule

5.4. Discussion

We have developed a high performing lung nodule classification approach using radiomic features of the lung and surrounding parenchyma extracted from CT data, and validated the performance in an independent validation cohort. We discovered that features from three separate perinodular parenchymal quartile-bands contributed various texture features to improve the model performance, at a level that was not achievable with one inclusive area of comparable size.

Other studies have explored the inclusion of perinodular features from the surrounding parenchyma for classification of lung nodules. A recent study comparing the performance of human observers to a computer algorithm showed observer interpreted broader characteristics such as spiculation, and disruption of perinodular parenchymal architecture as significant indicators of malignancy; however, subjective assessment of these characteristics is associated with high degree of observer variability[86]. Dilger et al (prior approach), demonstrated the potential of quantitative texture features for improved classification in a cohort of 50 subjects, using bounding boxes for capturing parenchymal signal approximately proportional to nodule size and whole lung density measures, with optimal classifier AUC-ROC of 0.938[10]. Huang et al. more recently demonstrated using a cohort of 186 subjects from the NLST trial that a machine learning system constructed with perinodular features achieved an AUC-ROC of 0.915[24]. Our comparison with this tool shows a direct overlap of two nodule selected features: nodule entropy and nodule variance. Also, two of their perinodular features selected from the small parenchyma ring surrounding the nodule were similar to our selected parenchyma quartile-band features: surrounding variance (at 25%, 50%) and surrounding parenchyma maximum intensity (at 25%, 50%, 75%).

The method of feature set selection used in our study is not only independent of classifier performance but also provides separate insight into the connections amongst imaging features and between characteristics and disease classification. In this study, the top two features were extracted from parenchymal bands distant to the nodule which provides evidence that there are more global changes in the lobe characteristics that imaging can detect. Decoding the spatial relationship between radiomic features from the parenchyma surrounding lung nodules presents future opportunity to advance the QIC-RATE tool analysis beyond a binary diagnosis. The field of transport oncophysics is relatively new, but holds promise in understanding the mass transport differentials of malignancy[99]. With a dataset classified in these differentials, the Extended QIC-RATE tool could be used as an effective delineator of mass transport.

This study did include some limitations. The malignant tumors in both the development and validation cohorts were larger on average than their benign counterparts creating a size bias between the classes. While RECIST diameter was selected in the final model (39/50), it was not predominantly

ranked, the other selected features were not highly correlated with the nodule size. While this bias exists in both cohorts, there was still a range in nodule size with some small malignant cases and some large benign cases; if size was a driving factor, we would have expected to see a greater disparity in performance in these nodules particularly. The CT data quality used in this study is not the current clinical standard (LDCT or clinical chest with contrast) but rather a cohort of high-resolution multi-center trial CT scans; however, it does demonstrate the performance advantage in using high quality scans and incorporating the perinodular signal. Our group has previously demonstrated the effects of LDCT and ultra-LDCT protocols on quantitative lung and airway measurements[100]. The model here purports the diagnostic quality of features extracted from high quality scans, of which most were not LDCT scans or subjects eligible for LDCT screening. Further studies investigating the transference of these high-resolution features to LDCT is needed to determine the performance of the Extended tool on lower-resolution scans. If transference of features to LDCT were to decrease the performance of the QIC-RATE tool, this would show the increased value of high-dose CT for the characterization of disease and reduction of repeated imaging studies would keep radiation dosage low.

We included only solid nodules in this study. In our validation cohort, only 34% would have met LDCT screening eligibility, making comparisons to the Fleischner guidelines more suitable for comparison than Lung-RADS. Assuming all follow-ups complied with the guidelines and patients were seen at the earliest follow-up point, the Extended QIC-RATE tool would reduce patient wait-time on malignant nodules by a cumulative 165 months, or on average 3.3 months per patient.

This study included a large dataset with histopathology confirmed malignant cases. The dataset we have assembled includes multi-center variability, indicative of generalizability to a wide study population. As the algorithm is based only on radiological features, the approach presents a pipeline integration advantage without the need for separate (and potentially subjective) data extraction and inclusion. The high accuracy of our approach can support clinician's higher confidence in risk assessment output and hence adherence to follow-up in concordance with the assigned class. This presents the potential to decrease the burden of un-necessary clinical follow-up of benign tumors and the timely and efficient treatment of those with cancerous tumors.

This chapter demonstrated the proposed approach for the QIC-RATE classification tool pipeline development using methods that allow for feature-transparency. Here, we demonstrate the QIC-RATE tool's accuracy using nodule standardized, perinodular parenchyma features. We quantified the theoretical benefit the Extended QIC-RATE -tool could have on the follow-up response compared to the Fleischner Society Pulmonary Nodule Follow-up Guidelines in reducing follow-up of benign nodules and expediting treatment of malignant nodules. The high performance of this method lends credence to its ability to be applied to more complex classification problems. The remaining chapters of this dissertation

use the QIC-RATE pipeline to look at sub-classification of binary diagnosis into histoplasmosis vs non-small cell lung cancer (**Chapter 6: Application of QIC-RATE to Histoplasmosis Classification**), COPD-related risk variables and binary diagnosis of lung nodules (**Chapter 7: Application of QIC-RATE to Global Lung Measures**), and finally to demonstrate the transferability of this pipeline to other diseases, modalities, and body-parts we applied it to large cohort of breast mammography masses (**Chapter 8: Application of QIC-RATE to Breast Tumor Classification**).

CHAPTER 6: APPLICATION OF QIC-RATE TO HISTOPLASMOSIS CLASSIFICATION

6.1. Introduction

Histoplasmosis is a fungal infection which is endemic to Iowa and other regions of the Mississippi, Missouri and Ohio River vallies[101]. This disease often presents as a pulmonary nodule via radiographic imaging with x-ray or CT. Histoplasmosis contributes to the clinical problem of differentiating cancerous lung nodules from nodules of benign pathology on chest CT within endemic regions, particularly as the nature of histoplasmosis lends to increased avidity on positron emission tomography (PET). An investigation based on data collected from the NLST demonstrated that clinicians within 'histoplasmosis-belts' were more conservative (lower false-positive) with the assessment of solitary nodules than clinicians outside the endemic regions[102].

Traditional nodule tracking guidelines such as Fleischner follow-up may not apply to populations living in 'Histo-belts', where less aggressive interventions may be more appropriate[103]. Radiomic features could serve as an assistive method for further characterizing, beyond size, these tumors on the first imaging timepoint. There is little prior work on identifiable qualitative or quantitative imaging features that are indicative of a histoplasmosis nodule. This chapter provides a proof of concept on a matched cohort of clinical cases with confirmed histoplasmosis or non-small cell lung cancer (NSCLC) comparing human observer classification performance to QIC-RATE classification performance.

6.2. Materials and Methods

6.2.1. Study Population

Subject included were part of a larger cohort collected retrospectively from the University of Iowa Hospitals and Clinics, located in a region endemic for Histoplasmosis[11,104]. With Institutional Board Approval, radiology reports from thoracic CT scans were text searched for the terms "pulmonary nodule" or "lung nodule". The electronic medical records (Epic, WI) from identified patients were manually searched for inclusion criteria of having diagnosis of pulmonary nodule through histopathology and CT imaging of solitary pulmonary nodule (4-30mm) prior to diagnosis, further details about this dataset can be found in **Chapter 2**. These subjects were matched based on age, sex, and smoking history.

6.2.2. QIC-RATE Tool Application

The QIC-RATE pipeline described in **Chapter 5** was applied with slight modifications to the features to accommodate the high variability in CT acquisition protocol from the retrospective, clinically acquired data. To summarize, the nodule and surrounding parenchyma were segmented semi-automatically using a seed-click method described in **Chapter 4**. In a preliminary study of the full (unmatched) cohort, it was determined that the superior method of parenchymal inclusion for this diverse population was the

inclusive rings; utilization of the exclusive bands was performed as a comparison to rings with decreased performance (see **Appendix C.3** for details on parenchymal segmentation). The perinodular region identified was segmented into rings that were nodule size-standardized through a nodule mask dilation procedure at 25%, 50%, 75%, and 100% the diameter for five candidate tools (Nodule, Margin, Immediate, Extended, Extended+). One hundred and one quantitative imaging characteristics describing intensity and 2D texture were extracted from the nodule and perinodular regions; 17 QICs describing border, size, and shape features were also extracted from the nodule mask. Highly correlated QICs were clustered using k-medoids clustering and the resulting medoids were sent through information theory-based feature set selection. The selected feature set was used to build an ensemble of artificial neural networks (ENNs) to differentiate between Histoplasmosis and NSCLC using leave-one-subject-out cross validation (**Appendix A.3**) for performance measure assessment.

6.2.3. Observer Assessment

We performed a controlled observer study on the full cohort of 71 plus 29 repeated cases (total of 100 cases provided to the observer) to examine the inter- and intra- observer variability. Four observers (2 Radiologists, 2 Pulmonologists) of varying experience were each provided de-identified CT data and accompanying basic clinical information in a manner blinded to diagnosis. The clinical information provided included, subject age, sex, PET avidity, and radiology report noted presence of cavitation or calcification. The observers were asked to provide a categorical risk (low, med, high) for NSCLC and a continuous analog risk between 0 (likely histoplasmosis) and 1 (likely NSCLC).

6.2.4. Statistical Assessment and Performance Measures

Statistical performance was determined using the methods described in **Appendix A.** In brief, QIC-RATE and observer continuous analog risk assessment performance were measured using AUC-ROC (Delong) and Youden's J statistic. McNemar's test was used to compare binary classification differences. Interclass correlation coefficient (ICC) was used as the assessment of consistency or reproducibility of continuous (0-1) risk made by different observers on the same nodule, the guidelines put forth by Cicchetti were used for interpretation[105]. Weighted Cohen Kappa and percent agreement were used to assess the categorical agreement among readers[106].

6.3. Results

6.3.1. Matching Reduces Demographic and Size Bias in Cohort

A total of 151 suitable subjects (49 histoplasmosis, 102 NSCLC) were retrospectively identified from the University of Iowa. Cases were matched between Histoplasmosis and NSCLC based on subject: (1) sex, (2) age within ± 3-years, and (3) self-reported smoking history. As pack-years were significantly different between groups and the accurate collection of pack-year information is difficult in long-term

smokers, smoking history was split into three categories (1) never smokers, (2) < 30 pack-year history – not smoking eligible for low-dose CT screening, and (3) ≥ 30 pack-year history – smoking eligible for low-dose CT screening. This resulted in 71 unique subjects (31 histoplasmosis, 40 NSCLC) and 94 total matches (some subjects matched to more than one other subject). **Table 6.1** indicates demographical variables for the matched cohort. No statistical demographic difference was found between diagnosis groups and nodule size was not significant (p = 0.40).

Table 6.1: Demographics of matched cohort along with p-value comparison between histoplasmosis and NSCLC.

		Histoplasmosis	NSCLC	p
N		31	40	-
Sex		23F:8M	23F:17M	0.14
Age **Mean Range**		57 years ± 8, 46-76 years	59 years ± 7.6, 47-79 years	0.12
BMI **Mean, Range**		30.7 ± 8.2, 19.1-51.8	27.9 ± 8.2, 16.0-44.5	0.11
Pack-years **Mean Range**		14.5 ± 17.8, 0-66	26.5 ± 26.8, 0-92	0.03
Smoke Tertiary*	**Non**	10	7	0.10
	NSE	16	18	
	SE	5	15	
Location	**RUL**	9	20	0.051
	RML	4	2	
	RLL	11	4	
	LUL	5	10	
	LLL	2	4	
Nodule Diameter **Mean, Range**		11.4mm ± 3.5 4.7-20.6mm	12.1mm ± 3.3 4.4-17.1mm	0.40

Definition of abbreviations: NSCLC – non-small cell lung cancer; N – number of subjects; F – female; M – male; BMI – body mass index; Non – never smoker; NSE – less than 30 pack-year history; SE – 30 or greater pack-year history; RUL – right upper lobe; RML – right middle lobe; RLL – right lower lobe; LUL – left upper lobe; LLL – left lower lobe

6.3.2. IO Feature Set Selection Illustrates Features from the Parenchyma are Informative of Disease

Following a rule of thumb of one feature per five training subjects, a maximum of 14 features was allowed for development of the candidate tools. The four parenchymal inclusion QIC-RATE tools used between 9 and 12 features while the Nodule tool used 10 (**Table 6.2**). Neither RECIST diameter nor nodule volume was selected in any candidate tools but the water-equivalent diameter was selected in the Nodule and Margin. No single feature was selected in all five candidate tools. Parenchymal ring *Long Run High Gray-Level Emphasis* was selected in all candidates with perinodular signal inclusion. There was little overlap in the selected nodule vs parenchyma features (i.e. nodular *Low Gray-Level Zone Emphasis* and parenchymal ring *Low Gray-Level Zone Emphasis* were not both selected in the same tool) indicating that the signals detected from these regions are unique. In all, 42 of the total 51 features

selected in at least one of the candidate tools were textural with most candidate translational (included in multiple tools) being run-length or size-zone features.

Table 6.2: Selected QICs from among the five candidate QIC-RATE tools.

		QIC Name	Nodule	Margin	Immediate	Extended	Extended+
Nodule	BS	Mean Sphere Difference	4				
	SH	Sphericity	1				10
	SZ	H_2O Equivalent Diameter	3	8			
	IH	25th HU Percentile	9		7	7	
	LTEM	Mean-5	6				
		Kurtosis-16	8		10		3
	NGTD	Strength	10		3	9	8
	GLRL	Gray-Level Variance		1	1		
		High Gray-Level Run Emphasis	5	4	11	5	
	GLZS	Gray-Level Nonuniformity	2				7
		Low Gray-Level Zone Emphasis		2			
		Small Zone Emphasis	7				
		Zone-Size Non-uniformity		3	2		
		Small Zone High Gray-Level Emphasis		7	8		
Parenchyma	IH	Full-Width-at-Half-Maximum					5
	LTEM	Mean-1		11	9		
	NGTD	Busyness				1	9
		Contrast		9	4	6	2
	GLRL	Gray-Level Non-uniformity		12	6	8	6
		Low Gray-Level Run Emphasis				4	
		Long Run High Gray-Level Emphasis		5	5	3	1
	GLZS	Low Gray-Level Zone Emphasis				2	
		Small Zone High-Gray Level Emphasis		6			4
		Total	**10**	**12**	**11**	**9**	**10**

Definition of abbreviations: QIC – quantitative imaging characteristic; BS – border sphere comparison; SH – shape; Sz – size; IH – intensity histogram; LTEM – Law's texture energy measure; NGTD – neighborhood gray-tone difference; GLRL – gray-level run length; GLSZ – gray-level size zone

6.3.3. Tool Assessment Performance Improved with inclusion of Surrounding Parenchyma

The five candidate tools were run through feature-set selection and development on ENN using LOO; **Figure 6.1** demonstrates the range of predictions from the five candidate tools. The performance measures are summarized in **Table 6.3**. Pairwise Delong assessment showed no statistical difference between ROC curves (p-value between 0.12-0.99). Of the candidate tools, the Extended+ (incorporating parenchymal ring at 100% diameter) performed the best using LOO (AUC-ROC = 0.89) utilizing ten features, 4 from the nodule and 6 from the perinodular parenchyma. The top ranked feature was the parenchymal *Long Run High Gray-Level Emphasis*. Applying the Youden threshold, the Extended+ tool achieved 84% specificity and 83% sensitivity. Applying the 90%-sensitivity threshold (0.55) to the Extended+ tool achieved a specificity of 61%.

Table 6.3: Performance measures of candidate QIC-RATE tools applied using leave-one-subject out cross validation.

	AUC-ROC	Youden	Specificity	Sensitivity
Nodule	0.78	0.36	0.61	0.90
Margin	0.86	0.46	0.77	0.85
Immediate	0.79	0.38	0.65	0.80
Extended	0.88	0.41	0.81	0.88
Extended+	**0.89**	**0.60**	**0.84**	**0.83**

Definition of abbreviations: N – number of subjects; AUC-ROC – area-under-curve of receiver-operator characteristic; 90%-sens – threshold that achieves 90% sensitivity see Appendix A.1.3

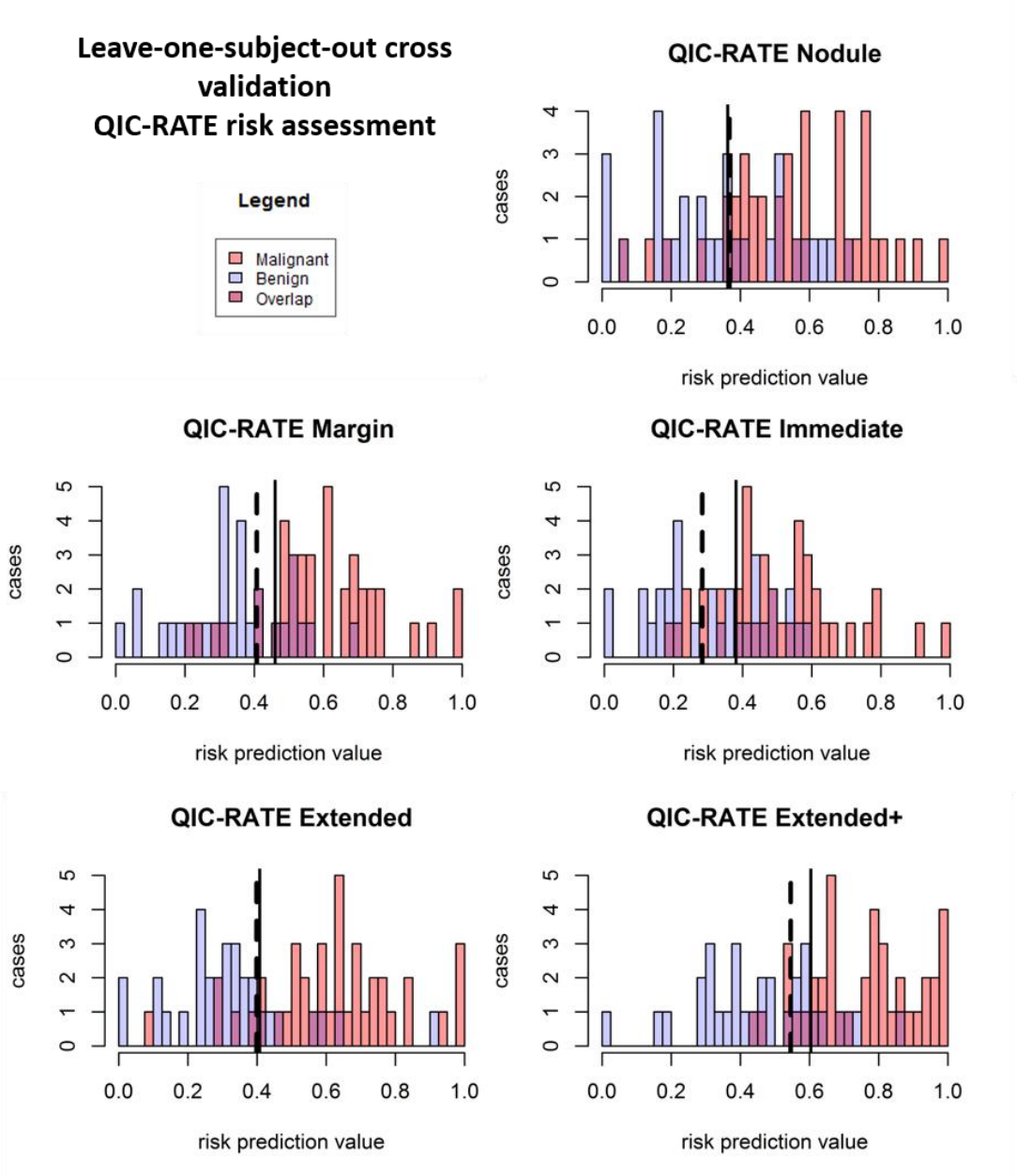

Figure 6.1: Overlay histogram visualization of five candidate QIC-RATE tools applied to the Histoplasmosis-NSCLC cohort. Solid lines indicate Youden threshold, dashed lines indicate threshold for 90% sensitivity.

6.3.4. Observer Categorical and Continuous Quantitative Assessments Demonstrate Variation Between Readers

Readers agreed on the categorical risk (low, medium or high) in 23 of the 71 cases; of those, 5 were scored low, 3 medium, and 15 high (**Figure 6.2**). All agreed-upon low-risk scored nodules were benign and all agreed upon high-risk nodules were NSCLC. This indicates that for 38% of the NSCLC cases there was high confidence in lung cancer classification among all observers and for 16% of the histoplasmosis cases there was high confidence in benign classification among all observers. Categorical assessment agreement was an average of 0.49 in weighted Cohen Kappa for all readers. The pulmonologists had the highest level of agreement between them (0.62) and the radiologists had the lowest level of agreement (0.36), however, some of the 'disagreement' could be the due to the risk-aversity of readers (i.e. one radiologist decided on an 'extreme' category – low/high while the other chose medium). In fact, categorical percentage of agreement hovered at random chance 32.4% –given there are three categories there is the unbiased draw likelihood that any rater will agree with another 33% of the time. On the quantitative assessment of risk, the ICC was 0.52 between all four raters indicating a fair level of agreement between raters. While differences existed between readers, assessment of the intra-reader differences in categorical assessments showed readers were 100% repeatable in category assignment. Continuous quantitative assessment was slightly less repeatable with a range in difference between <0.01 and 0.13.

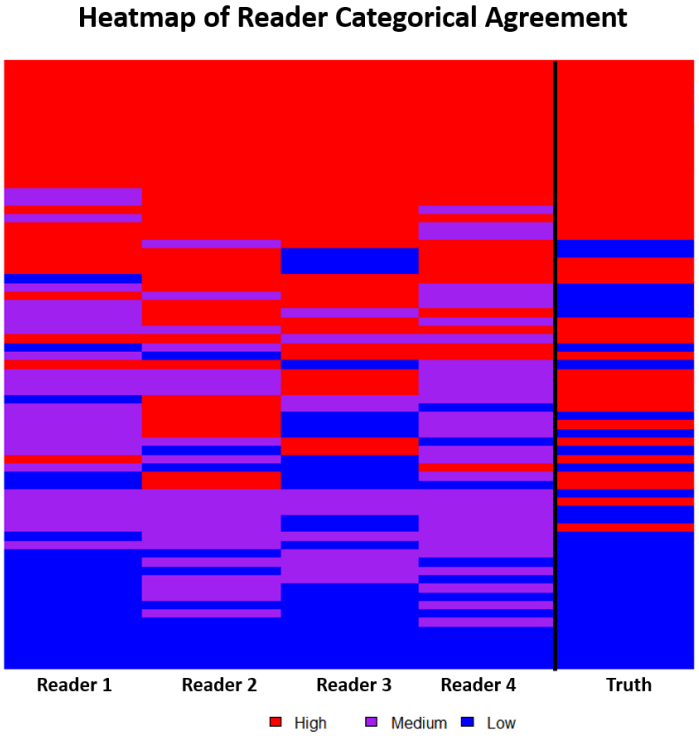

Figure 6.2: Heatmap of categorical agreement among readers. Colors: Blue - low risk; Purple - medium risk; Red - high risk.

6.3.5. Observer Continuous Quantitative Assessment Demonstrates Benefit of Human Observer, Potential for Simple Self-trained Tool

The observers' continuous quantitative risk scores (between 0-1) were assessed for direct comparison to the QIC-RATE tool. Observers ranged in AUC-roc = 0.65-0.80 (power 0.54-0.99) and Youden-threshold based sensitivity and specificity between 0.65-0.94 and 0.31-0.88 respectively (**Table 6.5**, **Figure 6.3**). Youden-threshold based sensitivity indicates the potential aid a simple 'self-training' tool could be assistive with compared to the categorical threshold. When compared to the average performance across all four observer readings, the QIC-RATE Extended+ tool had comparable sensitivity with an improved specificity.

Table 6.4: Performance Measures of quantitative (analog) risk assessment from the four human readers.

	AUC-ROC	Youden	Specificity	Sensitivity
Reader 1	0.76	0.63	0.62	0.88
Reader 2	0.80	0.67	0.73	0.79
Reader 3	0.74	0.74	0.88	0.65
Reader 4	0.65	0.75	0.31	0.94
Average Reader	0.74	NA	0.63	0.82
QIC-RATE Extended+	0.89	0.61	0.84	0.83

Definition of abbreviations: AUC-ROC – area-under-curve of receiver-operator characteristic; NA – not applicable

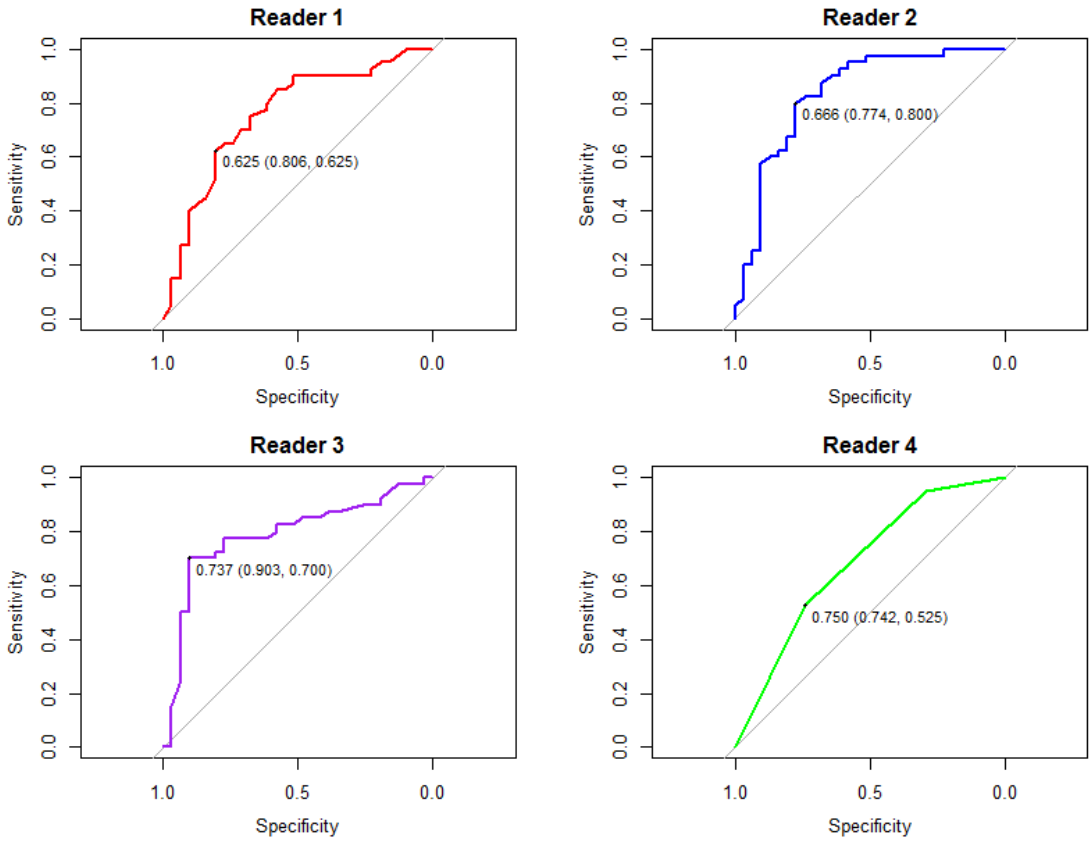

Figure 6.3: Receiver-operator characteristic curves for the four reader's continuous risk scores.

6.4. Discussion

In this chapter we have demonstrated the transferability of the QIC-RATE pipeline to a cohort of retrospectively collected clinical CT scans. This study was a proof of concept applying machine learning techniques developed and tuned on a large research cohort with malignant/benign distinction, to the problem of Histoplasmosis/NSCLC distinction; with consideration of the impact from surrounding perinodular region's signal. It further compared predictive results from the model to observers with significant experience in distinction of Histoplasmosis from other pulmonary nodule developing diseases. It confirmed the model developed on a small cohort and utilizing only CT extracted features, could achieve predictive performance that is comparable to a human reader.

Previously we have shown in **Chapter 5** that incorporation of perinodular signal significantly improves discriminatory ability on a large cohort of qCT scans with less variation in acquisition protocol. In the current chapter, the incorporation of perinodular signal also improved the performance at a level that approached significance between the Nodule and Extended/Extended+ tools. In the best performing tool, Extended+, six of the features were selected from the perinodular region including five texture and one intensity histogram measure. Several of those features – *Contrast*, *Gray-Level Non-uniformity in Runs*, *Long Run High Gray-Level Emphasis* – were selected in all four QIC-RATE incorporating perinodular QICs, indicating this textural signal is beneficial to the classification problem regardless of the size-standardization amount. Several features were selected in only the Extended+ QIC-RATE, including the perinodular *Small Zone High Gray-level Emphasis* and *Full-Width-at-Half-Maximum*, indicating the size-standardization for these features highlights their usefulness.

It is likely that increased performance could be achieved with this method (similar to **Chapter 5**) with CT protocol standardization. For example, in the retrospective clinical cases used in this investigation, subjects were not coached to a particular lung volume and differences in lung inflation likely reduce signal integrity of the perinodular features extracted[107]. Also, a large proportion (66/71) subjects in this cohort had iodine contrast-enhanced scans which can affect the extraction of derived measures[108]. As slice thickness was much larger in this cohort (mean = 3.30mm), we adapted the feature extraction pipeline for 2D textural features extracted from the slices containing nodule and perinodular region. It has been previously shown in the classification of lung cancer brain metastases that 3D textural features are more descriptive than 2D features[5]. The studies used here were acquired between October 2007 and December 2014; with increased technological advances making their way into the clinical setting, such as faster-acquisition LDCT and improved reconstruction algorithms, there will be improvements over time with the z-plane thickness which would make 3D features a more powerful option.

While the observers did differ in their categorical and continuous risk scores on a per nodule basis, they did perform well in distinguishing histoplasmosis to non-small cell lung cancer and their categorical risk scores were 100% repeatable on the intra-reader analysis. The experience level of these individuals is high in the given task while the QIC-RATE tool was built only using the looped LOO training cases (subjects = 70, per run) the number of true histoplasmosis and NSCLC cases the observers have been trained on is orders of magnitude larger. We did see potential improvement in observers when using a simple linear discriminant (Youden threshold of risk score) implying that it may improve observer (intra) repeatability and consistency/accuracy to use a continuous risk that has been tuned to their own level of 'risk percentage application' as opposed to categorical assessment.

This study contained limitations. First, it was a retrospective study collected from a clinical cohort, leading to a diversity in scanning protocols – including contrast enhancement- and selection biases. Second, the sample size collected was small (N = 71) and due to case-control matching study design, the clinical proportions of the two disease states were not maintained. In true clinical practice, it is unknown the actual rate of pulmonary histoplasmosis as often patients with pulmonary histoplasmosis nodules are not symptomatic and likely many are not definitively diagnosed as histoplasmosis. The histoplasmosis cases in this study were histopathologically diagnosed, such that only cases clinically warranting invasive procedure were included. In addition, the presentation of data with only a two-class outcome (histoplasmosis versus NSCLC) does not reflect clinical practice, in which multiple benign and malignant categories exist. Interpretation of the QIC-RATE and human readers performance should acknowledge the targeted approach of this study and not infer clinical practice performance from these results. Additionally, the small cohort size limits both the number of features we allowed each QIC-RATE tool to implement in ENN and increases the potential effects of scanning variability and inter-subject biological variance irrelated to pulmonary nodule pathology.

In this chapter we have demonstrated the transferability of the QIC-RATE pipeline to clinical-quality scans and maintained improved performance with size-standardized perinodular signal inclusion compared to utilization of features only from the nodule. We have further compared the developed tool to four blinded expert readers, demonstrating while there is room for improvement in the risk prediction accuracy of the Extended+ tool, it performs on par with clinician assessment. The next chapter, **Chapter 7 – Application of QIC-RATE to Whole Lung Measures**, investigates the transferability of the QIC-RATE pipeline to features extracted from the lung, lobe, and airways with comparison to a statistical multivariate model framework.

CHAPTER 7: APPLICATION OF QIC-RATE TO GLOBAL LUNG MEASURES

7.1. Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by obstructive lung function. Evaluated with medical imaging, COPD is heterogenous with varying presentations of structural changes in the lung parenchyma and airways. COPD is a risk factor for lung cancer development, independent of smoking history[41,109]. Prior studies have shown clinical (pulmonary function testing – PFT) and qualitative assessment links between COPD features and risk of lung cancer[41,110,111]. There has been limited published research into the overlap of the COPD-related qCT measures and risk of lung cancer in subjects with pulmonary nodules, with most studies including subjects with and without nodule presence[41,112-116]. Extent of emphysema in the lungs has been shown to be a positive predictor of lung cancer[112-114,116,117], qCT airway measures have also been assessed for potential predictive benefit however no significant discriminatory ability has been demonstrated[114,117]. However, many of these studies have not focused on controls with nodules – those individuals at a heightened risk of lung cancer by the sheer fact they have a pulmonary tumor.

This chapter investigates the utility of objectively and automatically obtained qCT metrics in predicting subjects with lung cancer on a cohort of scans all of which include pulmonary nodules ≥ 4mm. Here, the nodule was not extracted or segmented from the scans prior to qCT feature extraction. The purpose of this was to assess if pertinent qCT features could be obtained without pre-processing by human readers. Ultimately for a risk assessment pipeline to be most clinically helpful, there should be the requirement for as little human effort as possible. Radiologists have an already heavy workflow and therefore are unlikely to want to add additional (potentially helpful) assessments if it requires more time and effort. Here, the developed QIC-RATE system was compared to the least absolute shrinkage and selection operator (LASSO) regression analysis for feature set selection and classification performance on qCT features and demographical/clinical characteristics.

7.2. Materials and Methods

7.2.1. Study Population

The study cohort of was comprised of subjects retrospectively collected from three prospective research studies (COPDGene, INAHLE, and NLST) and included 327 individuals with pulmonary nodules (86 with primary lung cancer diagnosis) who underwent CT prior to diagnosis (see **Chapter 2** for additional information)[26,40,41]. The subjects were sectioned into a training cohort (n=278) and a validation cohort (n=49), using class-persevering random selection. Demographic and basic clinical features were

obtained from parent studies (**Table 7.1**). Inspiratory CT data was collected from multiple institutions following a standardized protocol.

Table 7.1: Subject demographics from the Development/Testing and Validation cohort.

			Malignant	Benign	p
Development/Testing	Subjects		71	207	-
	Age (mean±SD)		64.3±10.1	62.2±8.3	0.11
	Sex (Female: Male)		41:30	101:106	0.53
	Smoking History (Yes: No)		63:8	204:3	0.55
	Cancer History (Yes: No)		12:59	9:198	0.35
	Family Cancer History (Yes: No)		37:34	85:122	0.79
	Family Lung Cancer History (Yes: No)		18:53	68:139	0.59
	Cessation Time (mean±SD)		3.85±9.29	4.5±8.04	0.31
	Diameter, mm (mean±SD)		14.7±8.3	9.11±5	0.84
	Gold Stage	0	34	78	0.20
		1	8	44	
		2	18	42	
		3	7	32	
		4	4	10	
	FVC (mean±SD)		130.1±213.6	126.7±234.7	0.96
	FEV1 (mean±SD)		80.2±22.1	74.2±18.5	0.30
	FEV1/FVC (mean±SD)		1.05±0.54	1.24±0.67	0.33
Validation	Subjects		15	34	-
	Age (mean±SD)		60.36±7.65	61.7 ± 9.7	0.79
	Sex (Female: Male)		10:5	21:13	0.15
	Smoking History (Yes: No)		14:1	33:1	0.32
	Cancer History (Yes: No)		3:12	4:30	0.46
	Family Cancer History (Yes: No)		9:6	16:18	0.58
	Family Lung Cancer History (Yes: No)		6:9	12:22	0.5
	Cessation Time (mean±SD)		2.13±5.25	3.57±6.1	0.74
	Diameter, mm (mean±SD)		15.13±8.1	10.4±6.3	0.29
	Gold Stage	0	10	14	0.15
		1	2	6	
		2	2	9	
		3	0	5	
		4	1	0	
	FVC (mean±SD)		75.7±24.9	71.1±24.9	0.21
	FEV1 (mean±SD)		84.5±19.5	82.3±19.2	0.32
	FEV1/FVC (mean±SD)		1.22±0.45	1.3±0.53	0.55

Definition of abbreviations – SD – standard deviation; FVC – forced vital capacity; FEV1 – forced expiratory volume at 1 second.

7.2.2. Feature Groups

Two feature groups were collected and analyzed for predictive capabilities: Clinical and Imaging/QICs. Clinical features required input from a human based on recollection of the patient, clinical testing, or image reader assessment. Imaging qCT features were automatically extracted from the CT datasets using Apollo software suite (Vida Diagnostics, Coralville, IA).

7.2.2.1. Clinical features

A subset of this study's cohort was previously used to investigate the utility and consistency of post-imaging mathematical prediction models for the differentiation between malignant and benign lung nodules (**Chapter 3**).The clinical predictive values collected were included as clinical features. Seven measures of subject-provided historical information, 1 measures of radiologist reported information, and 4 PFT.

7.2.2.2. Extraction of Quantitative Imaging Characteristics

In total, 183 qCT measures were available for model development. QCT characteristics of the parenchyma (whole lung and lobar) and airways (segmental branches) were extracted. These included measures of image intensity from the lung tissue (mean, standard deviation, skewness, kurtosis, etc.), airway characteristics (wall thickness, lumen areas, pi10, etc.), volume characteristics, and low-attenuation area percentiles. For this study, all available qCT measures from the Apollo reports were used. The coefficient of variation (CV) among lobes and airway paths of the features was calculated using the 'raster' package in R[70].

7.2.3. Application of Statistical and Machine Learning Techniques

Model and feature performance were tested using methods described in the **Appendix A** including AUC-ROC (Delong) and Youden J statistic. The models were developed using two feature pools, qCT (imaging features only) and qCT+Clinical (qCT features alongside the clinical features).

7.2.3.1. Univariate Analysis

Univariate statistical assessment was performed with associations to diameter assessed. Logistic regression was utilized to assess the association between imaging parameters and malignancy status. Estimated effects of predictors are reported with odds ratios (OR) scaled to one standard deviation change. Imaging parameter performance was evaluated using the c-statistic (AUC-ROC). All statistical testing was two-sided and assessed for significance at the 5% level using SAS v9.4 (SAS Institute, Cary, NC).

7.2.3.2. Multivariate Model Development

Using the training dataset, LASSO models were applied to identify prognostic predictors of nodule malignancy status. The LASSO penalty parameter model performance metrics were derived from 100 iterations of 10-fold cross-validation. Observed and optimism-adjusted AUC, calibration plots, and confusion matrices are reported. The model derived in the building phase was applied to the testing

dataset. The performance of the model was assessed by AUC-ROC, reported is mean optimize adjusted AUC-ROC.

7.2.3.3. QIC-RATE Model Application

The QIC-RATE pipeline described in **Chapter 5** was implemented with the slight modifications. Image segmentation and feature extraction was implemented using the Apollo software. Highly correlated features are reduced to a single representative feature through k-medoid clustering, the reduced feature set undergoes the IO set selection method to obtain a ranking of informative predictors, and the selected feature set is used to train ENN. Here we apply the techniques developed for feature set reduction, selection, and classification to the training set on feature groups qCT and qCT+Clinical. The final trained models (development) were applied to the testing cohort.

7.3. Results

7.3.1. Statistical and Machine Learning Techniques Results

The set selection methods (Multivariate and IO) were applied to the feature pools qCT and qCT+Clinical, **Table 7.2** indicates the features selected in each of the models. In total, 32 features were selected as predictors in a model. There was minimal overlap between the features selected by Multivariate and IO methods (2 features overlap – lobe percent above 0 HU and diameter). **Table 7.3** shows the performance of the feature set selection methods (Multivariate, IO) and classification methods (LASSO, ENN). The univariate analysis showed the top three predictors to be intensity histogram based of the lobe with the nodule (*histstandard_deviation*, *skewness*, *kurtosis*) with training AUC-ROC between 0.68 and 0.71 ($p < 0.01$). Controlling for nodule size (diameter) and COPD Gold Stage did not affect the training AUC-ROC by more than 0.01 on any univariate model.

7.3.1.1. Multivariate Analysis selects Diameter and qCT features for highest training performance

The multivariate analysis yielded a model that incorporated QICs from the airway tree, whole lung, and lobe. Selecting only automatically extracted imaging features for model development included 7 measures, with a training AUC-ROC of 0.75 and a testing AUC-ROC of 0.56 – indicating overtraining on the training dataset. Allowing the model to select clinically ascribed features from radiologist or subject input produced a model included diameter and four qCT features, with a training AUC-ROC of 0.77 and testing AUC-ROC of 0.62; the improvement in validation AUC-ROC could point to size bias within our cohort. A model developed using only clinical/demographical features included only the diameter, with a training AUC-ROC of 0.70 and validation AUC-ROC of 0.64.

Table 7.2: Selected features for each of the developed models with odds ratio.

Category- Location	Feature	OR	M-I	IO-I	B-CI	M-CI	IO-CI	M-C
Clinical	Diameter	-			X	X	X	X
Histogram-Lung	standard_deviation	2.52	X		X	X		
	totalVolcm3	0.73		X			X	
Histogram-Lobe	percent_above_0	1.47	X	X			X	
	skewness	0.49		X			X	
	histmean	1.67		X			X	
	percentBelow910	0.75		X				
Airway-Lung	stdDevMaxWallThickness	1.06	X			X		
	stdDevAvgWallThickness	0.67	X			X		
	pi10_leq	1.51		X			X	
Airway-Lobe	avgWallAreaFraction	0.61		X			X	
Airway-AP	avgWallAreaFraction	1.05		X			X	
	avgMinWallThickness	0.79		X			X	
	avgMinorInnerDiam	1.23		X			X	
	stdDevMaxWallThickness	0.52	X			X		
	stdDevAvgWallThickness	1.16	X					
CV-Lobe	stdDevMajorOuterDiam	0.83	X					
	avgInnerPerimeter	0.94		X			X	
	stdDevAvgWallThickness	0.98		X			X	
	stdDevMinWallThickness	0.82		X				
	avgInnerEquivalentCircleDiam	0.94		X			X	
CV-AP	avgMaxWallThickness	1.29		X			X	
	avgOuterArea	0.93		X			X	
	avgMinorOuterDiam	1.09		X			X	
		Count	**7**	**17**	**2**	**5**	**16**	**1**

Definition of abbreviations: OR – odds ratio; CV – coefficient of variation; B-CI – bi-variate model qCT+clinical; M-I – multivariate selection qCT; IO-I – information optimization selection qCT; M-CI – multivariate selection qCT+clinical; IO-CI – information optimization selection qCT+clinical; M-C – multivariate selection clinical

Table 7.3: Performance results from the developed models using QICs and/or clinical characteristics.

Feature Groups	Selection	Model	Training AUC-ROC	Testing AUC-ROC
qCT	Multivariate	LASSO	0.75	0.56
	Multivariate	ENN	0.81	0.60
	IO	ENN	0.74	0.74
qCT + Clinical	Multivariate	LASSO	0.77	0.62
	Multivariate	ENN	0.81	0.62
	IO	ENN	0.77	0.79
Clinical	Multivariate	LASSO	0.70	0.64

Definition of abbreviations: AUC-ROC – area-under-curve of receiver-operator characteristic; qCT – quantitative computed tomography; LASSO - least absolute shrinkage and selection operator regression analysis; ENN – ensemble of artificial neural networks; IO – information optimization

7.3.1.2. ENN Classification Schema Improves Testing Performance of Multivariate Selected Imaging Features

The features selected through multivariate model development were used to train ENN models. The resulting training performance was similar to the multivariate model training performance, however the resulting testing performance from the ENN model was higher (0.60) for the qCT-only features model

compared to the multivariate testing (0.56). This indicates the ENN classification would be potentially more useful on new cases.

7.3.1.3. Combination Medoids-IO and ENN Less Likely to Over-train than Multivariate Approach

While the multivariate selection method for features obtained higher training AUCs for both Feature Pools (qCT+Clinical and qCT-only), the testing AUCs for these models implied potential overfitting, particularly with the qCT-only features. This overfitting persisted in the ENN trained with the Multivariate selected features indicating the overfitting is occurring during feature set selection. In contrast, the Medoids-IO selection followed by ENN model development obtained testing AUCs that more consistent with the training AUCs. Delong's analysis of the testing AUCs demonstrated that the qCT+Clinical Medoids-IO selected ENN model was significantly better on new cases than both the Multivariate selected ENN models. The qCT-only IO selected ENN model was not statistically better but was still higher in AUC.

7.3.2. Quantitative Imaging Feature Importance

7.3.2.1. Nodule Diameter only Clinical Feature selected by both Multivariate and IO

The nodule diameter, measured as the RECIST diameter, was the only clinical characteristic selected by the models. This could be due to the size bias that is seen in the cohort of pulmonary nodules – with cancerous nodules tending on average to be larger than their benign counterparts. A logistic regression model using diameter achieved a training AUC-ROC of 0.70 and validation AUC-ROC of 0.64. Diameter was selected in both Multivariate and IO selection methods and in the bi-variate model. The addition of clinical characteristics did not significantly improve the performance of either Multivariate or IO set selection in ENN development ($p > 0.05$). The addition of clinical characteristics in IO set selection did significantly improve the testing AUC-ROC over Multivariate set selection without clinical features ($p = 0.02$).

7.3.2.2. K-medoids Clustering Interrogates Feature Correlations

For the full training dataset, the optimal k produced 37 clusters. Example clustering of QICs and clinical features is shown in **Figure 7.1**. Across 10-fold kCV, clustering of the QICs using k-medoids method showed that 21 were stable (medoids in all 10 folds) and an additional 6 (27 total) which were semi-stable (medoids in at least 8 of the 10 folds). Of note, the lobe *standard_deviation* QIC which was selected as the bi-variate and in both the multivariate models was not selected as a medoid in any of the 10-folds; instead it was either a member of the cluster with representative lobe mean intensity (8 folds) or

lobe skewness (2 folds). As the IO feature selection is performed only on medoids, this feature was not available for selection using the IO method.

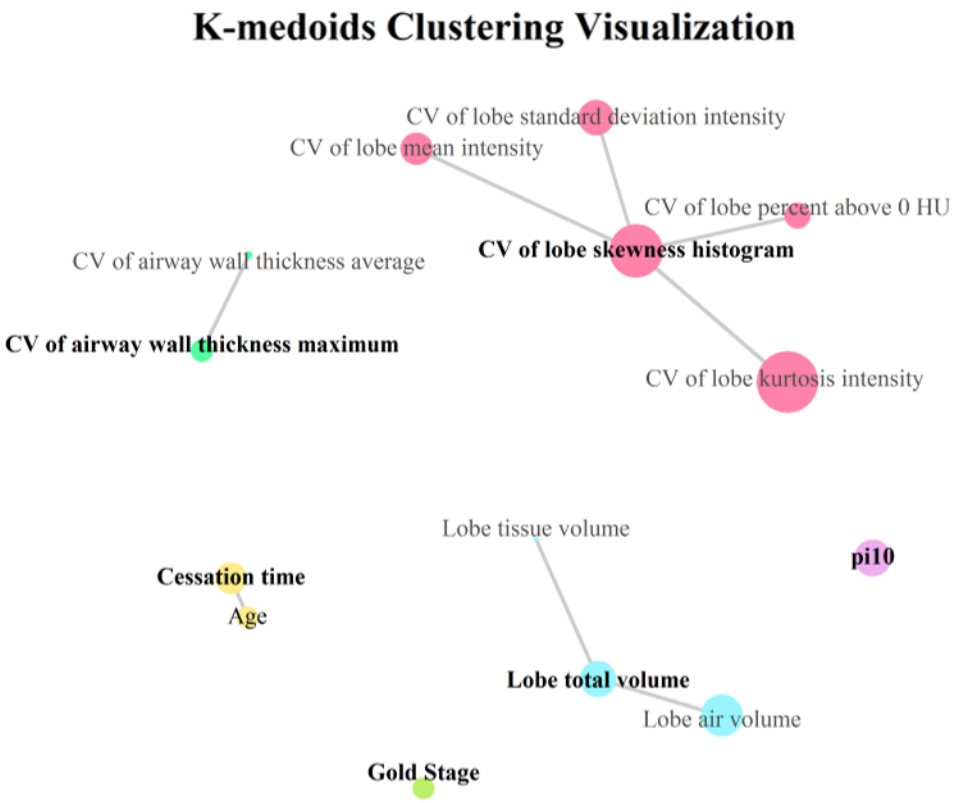

Figure 7.1: Example of clustering arrangement for select medoid features (bolded). Clusters are color coded. Lines indicate strength of correlation between features within a cluster. The size of the feature point indicates the information theory metric (larger circle means the feature shares more information with diagnosis). Definition of abbreviations: CV – coefficient of variation

7.3.2.3. Multivariate and IO select vastly different QIC features

Comparing the features selected by the two methods, only nodule diameter and the lobe percent above 0 HU were common (**Table 7.2**). This is likely largely due to the feature-set reduction by k-medoids performed prior to IO-set selection. Of the features selected by the multivariate model, only two (diameter and lobe percent above 0 HU) were selected as a medoid during 10-fold kCV k-medoids clustering. As such, none of the remaining multivariate selected features were available for IO-set selection. The IO method selected more CV features than the Multivariate selection method which only selected the CV of the Standard Deviation of the Major Outer Diameter using the qCT feature pool. The selected CV features were predominately from the airways.

7.4. Discussion

This study has demonstrated the potential richness in extra-nodular, automatically extracted imaging-derived features for the distinction between subjects with malignant and benign pulmonary nodules. It has highlighted the utility of more advanced methods of qCT feature selection for less overtraining. Here, we extracted a large number of features as an exploratory manner instead of selecting specific features as has predominately been reported previously. As such we have found that the QIC-RATE system with IO was advantageous for the exploratory manner of feature selection.

Prior works with qCT have primarily focused on the associations with lung cancer, irrespective of pulmonary nodule presence. These works have reported mixed results in the benefit of qCT for lung cancer risk assessment. Studies by Chubachi et al. and Gagnet et al. indicated that increased low attenuation areas percentage, indicative of CT characterized emphysema, were higher in subjects who developed lung cancer[112,113]. Studies by Gierada et al. and Wille et al. did not find statistically significant differences but indicated that emphysema was more frequently seen and at a higher grade in subjects with lung cancer[114,115]. Bae et al. investigated whole lung and lobar qCT emphysema ratios, finding the odds of lung cancer increased in lobes with more severe emphysema[116]. However, work from Wilson, Maldonado, and Johannessen showed no statistical evidence in quantitative lung parenchyma and/or airway measures and risk of lung cancer[110,111,117]. Schwartz et al found in multivariate modeling only the expiratory qCT measure -856 HU and PFT characteristics were independent predictors of lung cancer risk[41].

It was surprising that diameter was the only clinical features selected, particularly as the features included have been utilized in previously published lung cancer prediction models. However, with early lung nodule detection via CT based lung cancer screening size bias between malignant and benign nodules is expected to decrease (compared to incidentally discovered nodules); this could lead to the other contributing demographical and clinical factors having more importance. Also, the increased use of lung cancer screening and the associated mechanisms for structured reporting of patient data could standardized these factors further, allowing for potentially more useful information than is currently gathered.

The limitations of this study included the retrospective collection, with a focus only on solid pulmonary nodules. The cohort used in this study has a size bias between malignant and benign classes, and the nodule was not excluded (segmented) from the analysis. However, the difference between the average malignant nodule diameter (mean 14.7mm) and benign nodule diameter (mean 9.11mm) is considered very small when placed in the context of whole lung structure assessment. We only included inspiratory scans in the analysis, there are known expiratory scan measures (% below -856) which could be further assistive in the differentiation between malignant and benign cases. In this chapter we just investigated automated qCT lung features, taking into consideration the lobe location of lung tumor,

extracted using a proprietary software suite. Nodule volumetric segmentation can be challenging and has the potential to effect nodule specific feature extraction. Also, from a clinical workflow standpoint, a fully automated tool which does not require human interaction for nodule identification and segmentation would be advantageous.

This chapter concludes the dissertation work on pulmonary nodules which has spanned from previously published MPMs to image feature analysis of features correlated with COPD. The following chapter, **Chapter 8: Breast Tumor Classification**, demonstrates the flexibility and adaptability of the developed QIC-RATE pipeline on breast mammography data.

CHAPTER 8: APPLICATION OF QIC-RATE TO BREAST TUMOR CLASSIFICATION

This chapter is adapted from the paper, "Information Theory Optimization Based Feature Selection in Breast Mammography Lesion Classification", published in *Conf Proc IEEE Eng Med Biol Soc*[97].

8.1. Introduction

Cancer of the breast is the most common cancer diagnosis for women in the United States, resulting in an estimated 268,600 new cases and 41,760 deaths in 2019[4]. Annual breast cancer screening with mammography is recommended by the American Cancer Society for average-risk women starting at the age of 40[118]. The Breast Imaging Reporting and Data System (BI-RADS) was developed to standardize screening reporting among medical professionals and communicate with patients about cancer risk[35,119]. This includes assessment categories assigned by a radiologist after interpretation of the mammogram, providing a malignancy risk classification for encountered tumors. While this provides a structured framework for medical professionals, there is still room for improvement in the diagnostic capabilities and reduction of false-positives False-positive tumors in screening can lead to unnecessary follow-up procedures and stresses on the patient. Several imaging-based tools have been developed for the diagnosis of breast cancer in mammography screening exams using radiomic features extracted from mammograms[22,120-129]. We propose to build upon previously published machine learning methods by controlled examination of the impact of incorporating radiomic features from the peritumoral space surrounding a breast mass.

The goal of this chapter is an exploratory look at the utility of the described method beyond CT and beyond the lung and to compare the applied results to other published applications on a common cohort. Further, we provide insight into the relationship between quantitative features automatically extracted by the computer from the mammography data versus BI-RADS criterion as scored by radiologist. Through this study we demonstrate high diagnostic performance via a machine learning method which may provide a separate and augmentative risk assessment to that of radiologist interpretation with BI-RADS.

8.2. Materials and Methods

8.2.1. Study Cohorts

This study used 1115 mammographic images from the publicly available Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM)[43,130]. The CBIS-DDSM provides pathological diagnosis data, BI-RADS scoring (subtlety, breast density, and assessment), and mass segmentations. Of the scans included in this study, 568 were malignant and 547 were benign leading to a relative balance between the classes. This cohort was split into a development set (N=1000; 507

malignant, 493 benign) and a validation set (N = 115; 61 malignant, 54 benign). For more complete details on the origin datasets, please see **Chapter 2**.

8.2.2. Segmentation of Mass and Breast Parenchyma

The CBIS-DDSM included mass segmentations for every subject in this study. The mass segmentations were performed using a local level set framework based on the Chan-Vese model[131]. For this study, the whole breast was segmented using a combination of region growing from the mass border and Otsu thresholding. To ensure breast segmentation was performed adequately, 10% of the cohort (115 cases) were randomly selected and visually inspected by the author for completeness. For parenchyma ring feature extraction, the mass mask was grown using a binary image dilation to produce parenchyma quartile-rings: 25%, 50%, 75%, and 100% of the maximum diameter of the mass (**Figure 8.1**).

8.2.3. Application of QIC-RATE

The QIC-RATE tool developed in **Chapter 5** was adapted for use on breast mammography images. The following methods indicate the adjustments made to the QIC-RATE pipeline, a visual summary is depicted in **Figure 8.1**.

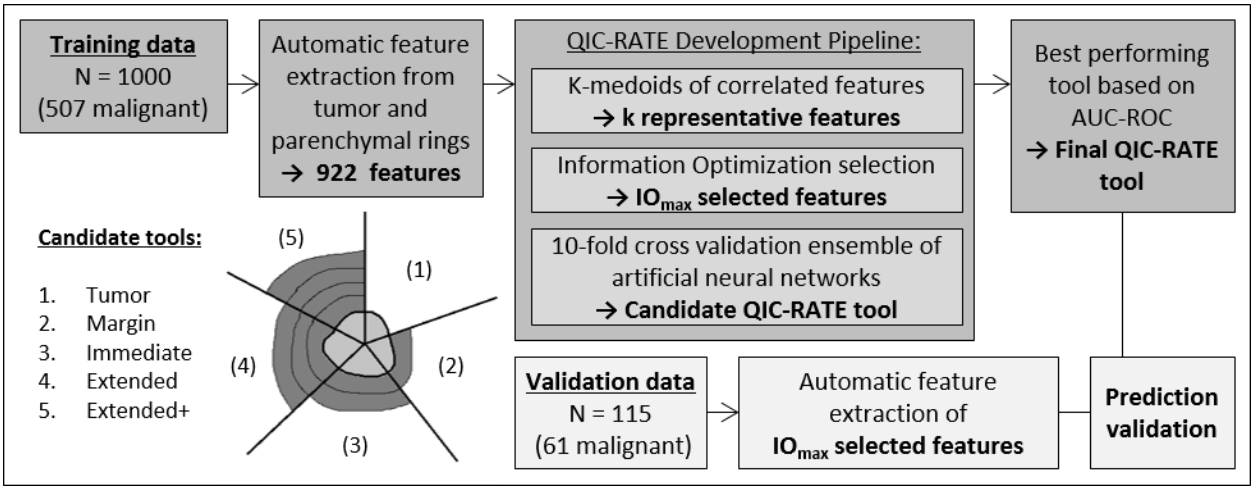

Figure 8.1: Overview of QIC-RATE tool development and validation pipeline for breast tumor application. Definition of abbreviations: IO – information optimization; AUC-ROC – area-under-curve of receiver-operator characteristic

Five candidate tools were developed including differing amounts of breast ring-parenchymal inclusion: [1] Tumor (no parenchyma), [2] Margin (tumor + 25% ring), [3] Immediate (tumor + 50% ring), [4] Extended (tumor + 75% ring), and [5] Extended+ (tumor + 100% ring). Two-dimensional quantitative features were automatically extracted from the mass and breast parenchyma quartile areas. Features from the tumor and parenchyma rings were applied to the QIC-RATE development pipeline for reduction (k-medoids clustering), selection (information optimization), and classifier training (10-kCV of ensemble of neural networks). The highest performing candidate tool was applied to the blinded

validation cohort for estimation of tool performance on novel cases. Finally, the best performing QIC-RATE tool was re-built to include BI-RADS categories to assess for the benefit of reader-determined features.

8.2.4. Performance and Comparison

Detailed information on the specific performance measures is included in the **Appendix A.** In brief, classification performance was assessed using AUC-ROC (Delong). The Youden's J statistic was used as the calibrated threshold with measures of sensitivity and specificity; also applied was the custom-threshold at training 90% sensitivity. McNemar's tests was used for statistical difference between binary classifications. Statistical comparison of feature values was performed for continuous and categorical variables. Comparison to BI-RADS categorization was achieved by splitting the reader assessment variable into high-risk (≥ Category 4) and low-risk (< Category 4); a Category 4 BI-RADS assessment is defined as having a suspicious abnormality and recommended follow-up is biopsy.

8.3. Results

8.3.1. Peri-Tumoral Signal Increases Performance

The highest performing candidate tool, the Margin QIC-RATE, was built using image features extracted from the tumor and the breast parenchyma within 25% of the tumor's radius (**Table 8.1**). Through 10-fold kCV, it achieved an AUC-ROC of 0.967 on the development cohort. Inclusion of three BI-RADS features improved the AUC-ROC to 0.968.

Table 8.1: Candidate QIC-RATE tool performance on breast mass classification in development dataset (10-fold kCV)

QIC-RATE tool	Features	AUC-ROC	Youden	Sensitivity	Specificity
Tumor	25	0.873	0.54	0.82	0.76
Margin	**38**	**0.967**	**0.56**	**0.86**	**0.95**
Immediate	43	0.942	0.57	0.84	0.92
Extended	63	0.943	0.57	0.84	0.92
Extended+	64	0.944	0.57	0.84	0.92
BI-RADS	3	0.857	0.52	0.79	0.75
Tumor & BI-RADS	28	0.925	0.63	0.84	0.90
Margin & BI-RADS	41	0.968	0.55	0.94	0.98

Definition of abbreviations: AUC-ROC – area-under-curve of receiver-operator characteristic; BI-RADS – Breast Imaging Reporting and Data System

Over one-third (15/38) of the Margin QIC-RATE selected IO_{max} were extracted from the surrounding breast parenchyma. Five of these features were IH describing both high and low order qualities of the parenchyma brightness profile. The values of these histogram qualities tended to be lower in malignant masses than in benign. The maximum intensity and histogram entropy were both statistically significantly different ($p < 0.01$) between malignant and benign cohorts, potentially demonstrating increased calcification and heterogeneity in the immediately surrounding parenchyma. Ten image texture

features were selected (6 LTEM, 3 GLRL, 1 GLSZ), including the feature selected first – Variance of LTEM 1 ($p < 0.01$). This feature is calculated from the principle components of features extracted from parenchyma filtered for ripple-spot textures. Twenty of the selected features came from the tumor region of interest (10 LTEM, 8 IH, 8 GLRL, 5 GLSZ, 2 SzSp, 2 NGTM). The remaining three features were extracted from the tumor's borders comparison with a circle of equal area.

8.3.2. Transparency in Features Allows for Analysis of Trends

Here we describe some trends in selected features between the tumor and the marginal parenchyma, for a complete list of the selected features see the supplementary material (**Table 8.2**). The spread of brightness values, described by the full-width-at-half-maximum of the intensity histogram, was statistically higher ($p = 0.023$) in malignant masses (0.014 ± 0.01) compared to benign masses (0.012 ± 0.01). Similarly, the full-width-at-half-maximum was higher in the marginal parenchyma of malignant tumors (0.012 ± 0.01) than in the marginal parenchyma of benign tumors (0.011 ± 0.01), although this trend was not statistically significant ($p = 0.128$). The GRLR gray-level non-uniformity demonstrates malignant tumors and the marginal parenchyma surrounding them were more homogenous in runs than their benign counterparts. The GRLR low gray-level run emphasis shows malignant tumors and their marginal parenchyma had less regions of low intensity than their benign counterparts.

8.3.3. QIC-RATE Tool Demonstrates Potential Increased Specificity Over BI-RADS

We compared the Margin QIC-RATE tool to a retrospective application of the BI-RADS assessment categories (**Table 8.3**). Tumors with BI-RADS assessment categories below 4 were designated as BI-RADS-benign and categories 4-5 as BI-RADS-malignant. Compared to the BI-RADS classification, the Margin tool had much improved specificity (94.9% compared to 54.2% from BI-RADS) with a slightly worse sensitivity (86.2% compared to 91.9% from BI-RADS). As QIC-RATE predicts risk assessment in the range of 0 and 1, we adjusted the risk threshold from 0.56 (Youden) to 0.47 (90% training sensitivity) as described in **A.1.3**. In the validation cohort, the Margin QIC-RATE with 90%-sensitivity threshold would have correctly predicted 6 more malignant tumors at the cost incorrectly labeling 3 benign tumors, compared to the unadjusted Margin tool (**Table 8.3**). In the validation cohort, the Margin QIC-RATE achieving sensitivity of 86.9% and specificity of 75.9% compared to BI-RADS assessment (93.4% and 61.1%).

Table 8.2: List of features selected in the Margin QIC-RATE tool.

#	Feature Name	Malignant		Benign		p
		Mean	SD	Mean	SD	
1	Parenchyma Variance of LTEM 1	-2.06E-1	3.30E+0	2.12E-1	3.30E+0	<0.01
2	Parenchyma Entropy	9.54E+0	8.56E-1	9.30E+0	8.56E-1	<0.01
3	Parenchyma Run-Length Gray-Level Non-uniformity	6.43E-2	2.21E-2	6.45E-2	2.21E-2	0.29
4	Tumor Kurtosis of LTEM 1	-1.19E-1	3.69E+0	1.22E-1	3.69E+0	0.39
5	Tumor Intensity 25th Percentile	2.91E+4	9.92E+3	2.70E+4	9.92E+3	<0.01
6	Tumor Intensity Mean	3.98E+4	8.40E+3	3.67E+4	8.40E+3	<0.01
7	Tumor Maximum In-plane Radius	2.63E+1	6.36E+0	2.40E+1	6.36E+0	<0.01
8	Tumor Circularity	5.39E+0	1.06E+0	5.25E+0	1.06E+0	0.14
9	Tumor Large-Zone Emphasis	3.65E+4	1.10E+5	1.71E+4	1.10E+5	<0.01
10	Tumor Zone-Percentage	1.37E-1	9.71E-2	1.56E-1	9.71E-2	<0.01
11	Parenchyma Intensity Variance	4.21E+7	5.93E+7	4.14E+7	5.93E+7	0.52
12	Tumor Variance of LTEM 1	-1.07E-1	3.69E+0	1.10E-1	3.69E+0	<0.01
13	Tumor Strength Texture	2.82E-1	2.29E-1	3.26E-1	2.29E-1	<0.01
14	Parenchyma Intensity Kurtosis	3.04E+0	2.06E+0	2.96E+0	2.06E+0	0.82
15	Parenchyma Low Gray-Level Run-Length Emphasis	1.22E-2	1.65E-2	1.50E-2	1.65E-2	0.20
16	Parenchyma Kurtosis of LTEM 1	7.85E-3	1.46E-1	-8.08E-3	1.46E-1	0.78
17	Tumor Low Gray-Level Run-Length Emphasis	6.40E-3	4.91E-3	6.94E-3	4.91E-3	0.46
18	Parenchyma Full-width-at-half-maximum	1.21E-2	8.35E-3	1.04E-2	8.35E-3	0.09
19	Tumor Busyness Texture	3.93E+0	2.86E+0	3.29E+0	2.86E+0	<0.01
20	Parenchyma Mean of LTEM 1	-5.14E-2	3.66E+0	5.29E-2	3.66E+0	<0.01
21	Tumor Long-Run High Gray-Level Emphasis	2.95E+3	4.89E+3	2.15E+3	4.89E+3	<0.01
22	Tumor Large-Zone Low Gray-Level Emphasis	1.63E+2	1.49E+3	1.35E+2	1.49E+3	<0.01
23	Tumor Long-Run Low Gray-Level Emphasis	8.51E-2	8.83E-1	4.74E-2	8.83E-1	0.05
24	Tumor Run-Length Gray-Level Non-uniformity	5.77E-2	1.57E-2	5.96E-2	1.57E-2	0.14
25	Tumor Full-width-at-half-maximum	1.46E-2	8.32E-3	1.23E-2	8.32E-3	0.02
26	Parenchyma Large-Zone High Gray-Level Emphasis	1.54E+7	4.06E+7	1.31E+7	4.06E+7	0.32
27	Parenchyma High Gray-Level Run Emphasis	2.66E+2	1.05E+2	2.65E+2	1.05E+2	0.81
28	Mean Absolute Border Comparison	4.91E+0	6.93E-1	4.66E+0	6.93E-1	<0.01
29	Tumor Long-Run Emphasis	6.18E+0	6.60E+0	5.35E+0	6.60E+0	<0.01
30	Parenchyma Intensity Maximum	4.85E+4	8.83E+3	4.66E+4	8.83E+3	<0.01
31	Parenchyma Skewness of LTEM 1	3.43E-3	2.84E-1	-3.52E-3	2.84E-1	0.59
32	Tumor Size-Zone Variance	2.96E-5	1.36E-4	4.12E-5	1.36E-4	<0.01
33	Tumor Mean of LTEM	2.93E-3	1.16E-1	-3.01E-3	1.16E-1	<0.01
34	Variance Absolute Border Comparison	6.92E-1	6.15E-1	6.89E-1	6.15E-1	0.58
35	Parenchyma Skewness of LTEM 2	2.05E-1	3.72E+0	-2.11E-1	3.72E+0	0.77
36	Parenchyma Mean of LTEM 2	6.82E-2	7.24E-1	-7.01E-2	7.24E-1	0.02
37	Kurtosis Absolute Border Comparison	2.63E+0	6.78E-1	2.61E+0	6.78E-1	0.50
38	Tumor Variance of LTEM 2	7.72E-3	5.19E-1	-7.94E-3	5.19E-1	0.90

Definition of abbreviations: SD – standard deviation; LTEM – Law's texture energy measures

8.3.4. BI-RADS Features Not Highly Correlated with Automatically Extracted Features

We can analyze the clustering of the feature set (including both QIC and BI-RADS features) to understand how computer-extracted features relate to those generated through radiologist assessment. Interestingly, the BI-RADs features were not clustered with any QICs (**Figure 8.2**). This indicates independent contribution of these features not directly captured through computer extracted radiomic

features. K-medoids analysis of the BI-RADS features showed similar trends regardless of parenchyma inclusions amount with all three being the medoid of their own cluster. The BI-RADS-*subtlety* cluster neighbored BI-RADS-*breast density* (range in separation). While breast density was more closely neighboring (separation: 0.625) a cluster composed of image texture features with medoid GLSZ zone percentage. The assessment cluster neighbored (separation: 0.795) a cluster with tumor and parenchyma IH entropy features. Subtlety was neighboring breast density (separation: 0.707).

Table 8.3: Contingency tables for comparison of retrospective application of BI-RADS assessment category.

		Binary Prediction	M	B	Sensitivity	Specificity
Development	BI-RADS assessment	Category ≥ 4	466	226	91.9%	54.2%
		Category < 4	41	267		
	Margin QIC-RATE Youden prediction	Malignant	437	25	86.2%	94.9%
		Benign	70	468		
	Margin QIC-RATE 90%-sens prediction	Malignant	466	57	91.9%	88.0%
		Benign	41	436		
Validation	BI-RADS assessment	Category ≥ 4	57	21	93.4%	61.1%
		Category < 4	4	33		
	Margin QIC-RATE Youden prediction	Malignant	47	10	77.0%	81.5%
		Benign	14	44		
	Margin QIC-RATE 90%-sens prediction	Malignant	53	13	86.9%	75.9%
		Benign	8	41		

Definition of abbreviations: M – true malignant; B – true benign; BI-RADS – Breast Imaging Reporting and Data System; 90%-sens – threshold derived from finding 90% sensitivity in development cohort (see A.1.3)

8.3.5. Performance Comparison to Other Published Approaches

The use of a publicly available dataset allows for comparison of different methods of classification in a more standardized way. As the CBIS-DDSM has been widely used for this classification task, the we are not able to provide a full comparison to all publicized applications. We have chosen instead to detail here recent applications to the CBIS-DDSM (**Table 8.4**), several systematic reviews are available that include applications that used a subset of the dataset[132-134]. Overall, our Margin QIC-RATE tool performed within the bounds set by AUC-ROC, accuracy, sensitivity, and specificity in those published articles. The best performing (based on AUC-ROC) of this group by Xie et al. was an approach using a combination of extreme learning machines and SVM of features (IH and texture) extracted from the tumor and parenchyma background regions[123].

Table 8.4: Recent publications incorporating the CBIS-DDSM cohort.

Publication	Cohort	AUC-ROC	Accuracy	Sensitivity	Specificity
Xie[123]	330	0.966	96.0	96.3	94.3
Abbas[135]	350	0.910	91.0	92.0	84.2
Verma[136]	200	NR	93.5	97.8	90.7
Jaffar[137]	1800	0.910	93.0	92.8	91.4
Zhang[124]	681	NR	84.4	NR	NR

Definition of abbreviations: AUC-ROC – area-under-curve of receiver-operator characteristic; NR – not reported

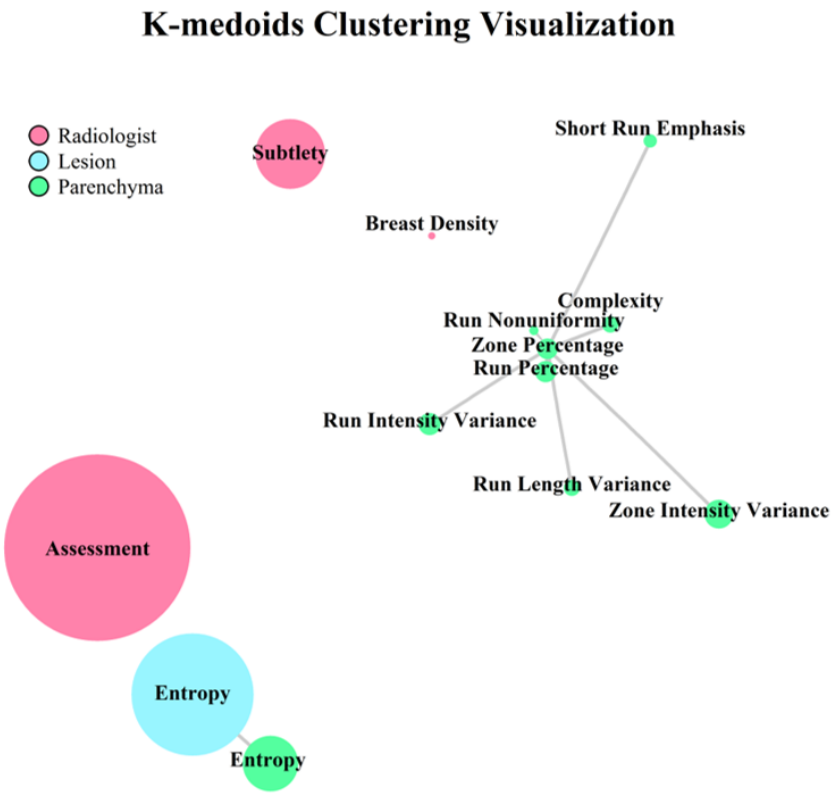

Figure 8.2: Visualization of the k-medoids clustering on the BI-RADS features (Radiologist) and their neighbor clusters with select medoid features (bolded). Lines indicate strength of correlation between features within a cluster. The size of the feature point indicates the information theory metric (larger circle - feature shares more information with diagnosis).

8.4. Discussion

We have demonstrated that the QIC-RATE pipeline, developed originally for the classification of lung nodules in CT, can be effectively applied with minor adjustments to other cancer classification problems. We have applied the developed QIC-RATE pipeline to a large, publicly available mammography dataset with confirmed mass diagnosis. In this chapter, we have demonstrated the potential benefit of including radiomic features extracted from the peri-tumoral breast tissue. Through a system of comparison tools, we have demonstrated a sufficient amount of tumor-standardized parenchyma to be 25% of the mass's diameter. Furthermore, we have compared the radiomic-based classification to the BI-RADS recommendations, showing the potential strength in the QIC-RATE tool in terms of specificity.

Other studies have also shown value in quantitative assessment of the breast parenchyma[22,123,125-129]. Zheng et al. utilized a lattice-based approach to extract features from regions of the breast parenchyma, with features (IH, GLRL, and co-occurrence) computed about the intersection points, the results of which were used in logistical regression classifier for breast cancer risk[125]. Sun et al. investigated features extracted from automated breast sub-regions segmented through 5-class fuzzy means clustering; the five sub-regions were ranked based on density value and fed into SVM classification

schema[126]. Li et al recently showed statistical improvement in a machine learning tool with the inclusion of normal contralateral breast parenchyma (skewness, Fourier power law beta)[127]. To the best of our knowledge this is the first study to standardize to the mass size the amount of surrounding breast parenchyma included. The tumors in this cohort varied in size from maximum-diameter 5mm to 67mm and breast size also varies among the population. Through this study we have examined the amount of surrounding parenchyma – by quartiles – and established that utilizing the signal from 25% of the tumor's maximal diameter is sufficient to extract meaningful radiomic characteristics pertinent to the distinction between malignant and benign breast masses.

We explored the potential clinical utility of employing the Margin QIC-RATE tool compared to BI-RADS classification. Using the Youden optimal threshold, QIC-RATE provided a higher specificity than using the BI-RADS assessment threshold of Category-4, however the sensitivity was lower. By adjusting the threshold to match the BI-RADS sensitivity in the development cohort improved the sensitivity in the validation cohort while still maintaining higher specificity than the BI-RADS. There is indication of some overtraining, shown by the lower performance of the QIC-RATE in the validation cohort. However, the maintained improvement in the validation cohort of both sensitivity and specificity over BI-RADS alone lends credence to the ability of a tool built solely on objective radiomic features to have a potential positive impact on clinical practice.

This study was limited in scope to 2D mammography images, there is a greater potential value of this approach (increased features, volumetric features) in 3D mammography images (tomosynthesis), however no large publicly available dataset has been compiled. Radiomic features can be sensitive to segmentation quality variability, in particular those describing tumor shape and border. The current approach utilized existing tumor segmentations, automated breast tumor detection and segmentation systems exist which could eventually allow this process to be streamlined for assessment without user requirements. Recently, risk stratification models have been promoted by the American Cancer Society to determine a woman's risk of breast cancer, and suggestions for increased screening (at 30-years of age) and increased used of other imaging modalities including MRI and ultrasound, at this time there are no large publicly available datasets with which to compare the utility of the described QIC-RATE approach on these methods. As the dataset was publicly available, some unknowns exist regarding data collection and processing; for example, as we saw in **Chapter 6,** variability often exists between readers and it is possible that the BI-RADS categories were collected from multiple readers. This could potentially influence linear correlation with automated features if the 'reader consistency' was low and could be a factor in why AUC-ROC was not improved with addition of BI-RADS features. Finally, BI-RADS does not fully capture all components of clinical risk that may be considered by a clinician in deciding the most

suitable follow-up for a patient, further studies in the observer-model performance are needed to address these questions.

This chapter has highlighted the flexibility of the QIC-RATE pipeline and performance validation for inclusion of surrounding tissues in solid tumors beyond the lung. The next two chapters draw together the conclusions of this thesis (**Chapter 10: Conclusions**) and highlight potential areas of further growth (**Chapter 9: Future Directions)**.

CHAPTER 9: FUTURE DIRECTIONS

As the modularity of the developed approach has proven useful, there is no shortage of possible future directions and applications. The following sections briefly notes potential areas of growth for QIC-RATE beyond the completed dissertation work.

9.1. Inclusion of Additional Imaging features

The developed method is neither tied to a number of features nor to a specific feature type – although the basis of the developed approaches presented in the thesis mainly focused on imaging characteristics. The science and engineering behind the QIC-RATE pipeline readily allows for the addition of expanded feature sets. New imaging-based features are being explored and developed for extraction in a multitude of applications, the addition of these features could highlight additional feature interaction and potentially increase information available for accurate diagnosis. Similarly, efforts are being made to standardize and ontologize imaging features for the comparison of different methods of calculating similar features (i.e. volumetric calculations based on pixel count or triangulation of surfaces)[80]

9.2. The Multiclass Approach

A natural next direction for the developed approach is the expansion from a binary class approach (malignant/benign, histoplasmosis/NSCLC) to an approach that delineates multiple classes simultaneously (adenocarcinoma/squamous cell carcinoma/histoplasmosis/tuberculosis/hamartoma). For this to be successful a large and protocol-controlled dataset is needed, with the increase in lung cancer screening and subsequent standardized follow-up on discovered nodules this is within reach of a large institution with a rigorous screening program.

9.3. Deep Learning

The arena of deep learning developed alongside this thesis work and has proven to be an efficient application for machine learning problems that traditional (non-deep) learning has encountered. The 'hands-off' black box can be an efficient and robust way of solving some time-intensive and methodology intensive challenges in image processing such as image registration and segmentation[138]. We believe deep learning and the QIC-RATE have a natural compatibility that when combined, can highlight the skills of both approaches.

Specifically, deep learning could be used to 'cut-out' the human interaction entirely by providing the nodule identification and image segmentation elements[139,140]. As nodule identification and segmentation does not require knowledge of true diagnosis, deep learning could tap into the large cohorts

with nodules identified and segmented – such as the LIDC – to train a tumor detection and segmentation method. From the deep learning identified segmentations, QIC-RATE could be employed on a cohort with known diagnosis to extract imaging features, determine the reduced and ranked set, and train the ENN. Given a large enough dataset – which could be soon possible with lung cancer screening – it would be interesting to also include a convolutional neural network as a 'second artificial reader' or an element in the ENN; thereby investigating if deep learning is pulling out the same information as the curated features or new measure that have not been curated yet.

CHAPTER 10: CONCLUSIONS

The results presented in this thesis support our hypothesis – *quantitative imaging characteristics (QICs) extracted from spatially-linked and size-standardized regions of surrounding tissue can improve risk assessment performance over features extracted from only the tumor regions*. Furthermore, we have met the goals to develop a flexible and robust pipeline for the extraction and selection of informative imaging-derived characteristics in medical imaging data. Using these characteristics, we have implemented a rigorous classifier development methodology which was validated both in training phase and through independent testing datasets.

The need for advanced methods of lung nodule risk stratification, post-CT identification, was highlighted through the investigation of existing MPMs, demonstrating need for improvement in specificity beyond calibration (Chapter 3). The primary objective for the developed QIC-RATE pipeline was the improvement of risk stratification of CT identified pulmonary nodules through standardized QICs, in which we were able to demonstrate exceptional performance including a validation accuracy of 98% (Chapters 4-5). A benefit of the developed system is the preservation of feature transparency throughout the reduction, selection, and classification process along with the ability to further explore the relationship of informative features. Further, the methodology was shown to be flexible to image and feature dimensionality as demonstrated in applications for breast tumor characterization (Chapter 8) and classification of histoplasmosis (Chapter 6). The methodology is adaptable to cohorts with heterogenous protocols; such as diverse CT acquisition protocols from retrospective clinical cases at UIHC (Chapter 6). In addition, the pipeline can easily incorporate additional features both imaging-related and non-imaging subject characteristics (Chapters 7-8). We have provided clinical context to the work by comparing theoretical nodule management procedures in accordance with clinical guidelines (i.e. Fleischner, Lung-RADS, BI-RADS) compared to risk stratification with the developed QIC-RATE tools.

With the growing use of medical imaging in the field of oncology, in lung cancer particularly, there is an increasing need for tools that provide diagnostic assessment with minimal added time, cost or risk to the patient. The work presented in this thesis addresses this need through establishing the QIC-RATE pipeline; a modular, scalable, transferrable pipeline for extracting, reducing and selecting, and training a classification tool based on QICs. Altogether, this resulted in a methodology that is validated, stable, high performing, adaptable, and transparent.

REFERENCES

1. Giger ML, Chan HP, Boone J: Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. Medical physics 2008, 35(12):5799-5820.
2. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A et al: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer (Oxford, England : 1990) 2012, 48(4):441-446.
3. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D et al: Radiomics: the process and the challenges. Magn Reson Imaging 2012, 30(9):1234-1248.
4. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2019. CA: a cancer journal for clinicians 2019, 69(1):7-34.
5. Ortiz-Ramon R, Larroza A, Ruiz-Espana S, Arana E, Moratal D: Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study. European radiology 2018, 28(11):4514-4523.
6. Dingemans KP, Mooi WJ: Invasion of lung tissue by bronchogenic squamous-cell carcinomas: interaction of tumor cells and lung parenchyma in the tumor periphery. International journal of cancer 1986, 37(1):11-19.
7. Sieren JC, Smith AR, Thiesse J, Namati E, Hoffman EA, Kline JN, McLennan G: Exploration of the volumetric composition of human lung cancer nodules in correlated histopathology and computed tomography. Lung cancer (Amsterdam, Netherlands) 2011, 74(1):61-68.
8. Sieren JC, Weydert J, Bell A, De Young B, Smith AR, Thiesse J, Namati E, McLennan G: An automated segmentation approach for highlighting the histological complexity of human lung cancer. Annals of biomedical engineering 2010, 38(12):3581-3591.
9. Sieren JC, Weydert J, Namati E, Thiesse J, Sieren JP, Reinhardt JM, Hoffman EA, McLennan G: A process model for direct correlation between computed tomography and histopathology application in lung cancer. Academic radiology 2010, 17(2):169-180.
10. Dilger SK, Uthoff J, Judisch A, Hammond E, Mott SL, Smith BJ, Newell JD, Jr., Hoffman EA, Sieren JC: Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. Journal of medical imaging (Bellingham, Wash) 2015, 2(4):041004.
11. Dilger SKN: Pushing the boundaries: feature extraction from the lung improves pulmonary nodule classification. Univeristy of Iowa; 2016.

12. Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, Bogot N, Zhou C: Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Medical physics 2006, 33(7):2323-2337.
13. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E: Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Medical physics 2009, 36(7):3086-3098.
14. Lee MC, Boroczky L, Sungur-Stasik K, Cann AD, Borczuk AC, Kawut SM, Powell CA: Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. Artificial intelligence in medicine 2010, 50(1):43-53.
15. Zhu Y, Tan Y, Hua Y, Wang M, Zhang G, Zhang J: Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. Journal of digital imaging 2010, 23(1):51-65.
16. Ferreira JR, Jr., Oliveira MC, de Azevedo-Marques PM: Characterization of Pulmonary Nodules Based on Features of Margin Sharpness and Texture. Journal of digital imaging 2018, 31(4):451-463.
17. Sun T, Zhang R, Wang J, Li X, Guo X: Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data. PloS one 2013, 8(5):e63559.
18. Amir GJ, Lehmann HP: After Detection: The Improved Accuracy of Lung Cancer Assessment Using Radiologic Computer-aided Diagnosis. Academic radiology 2016, 23(2):186-191.
19. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby RA, Balagurunathan Y et al: Predicting Malignant Nodules from Screening CT Scans. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2016, 11(12):2120-2128.
20. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM: Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Scientific reports 2016, 6:24454.
21. Nibali A, He Z, Wollersheim D: Pulmonary nodule classification with deep residual networks. International journal of computer assisted radiology and surgery 2017.
22. Sun W, Zheng B, Qian W: Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. Computers in biology and medicine 2017, 89:530-539.
23. Ciompi F, Chung K, van Riel SJ, Setio AAA, Gerke PK, Jacobs C, Scholten ET, Schaefer-Prokop C, Wille MMW, Marchiano A et al: Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific reports 2017, 7:46479.

24. Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, Hussien A, Rathmell J, Thomas B, Chen C et al: Added Value of Computer-aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study. Radiology 2017:162725.
25. Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N: A Combination of Shape and Texture Features for Classification of Pulmonary Nodules in Lung CT Images. Journal of digital imaging 2016, 29(4):466-475.
26. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine 2011, 365(5):395-409.
27. Wood DE: National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines for Lung Cancer Screening. Thoracic surgery clinics 2015, 25(2):185-197.
28. Jaklitsch MT, Jacobson FL, Austin JH, Field JK, Jett JR, Keshavjee S, MacMahon H, Mulshine JL, Munden RF, Salgia R et al: The American Association for Thoracic Surgery guidelines for lung cancer screening using low-dose computed tomography scans for lung cancer survivors and other high-risk groups. The Journal of thoracic and cardiovascular surgery 2012, 144(1):33-38.
29. Wender R, Fontham ET, Barrera E, Jr., Colditz GA, Church TR, Ettinger DS, Etzioni R, Flowers CR, Gazelle GS, Kelsey DK et al: American Cancer Society lung cancer screening guidelines. CA: a cancer journal for clinicians 2013, 63(2):107-117.
30. Wiener RS, Gould MK, Arenberg DA, Au DH, Fennig K, Lamb CR, Mazzone PJ, Midthun DE, Napoli M, Ost DE et al: An official American Thoracic Society/American College of Chest Physicians policy statement: implementation of low-dose computed tomography lung cancer screening programs in clinical practice. American journal of respiratory and critical care medicine 2015, 192(7):881-891.
31. Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, Franks K, Gleeson F, Graham R, Malhotra P et al: British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax 2015, 70 Suppl 2:ii1-ii54.
32. Blackmon SH, Feinglass SR: The United States Preventive Services Task Force recommendations for lung cancer screening. Thoracic surgery clinics 2015, 25(2):199-203.
33. Radiology ACo: Lung CT Screening Reporting and Data System (Lung-RADS). In: https://wwwacrorg/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads. vol. 2018: American College of Radiology; 2014.
34. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M et al: Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology 2017, 284(1):228-243.

35. Mercado CL: BI-RADS update. Radiologic clinics of North America 2014, 52(3):481-487.
36. Hammond E, Chan KS, Ames JC, Stoyles N, Sloan CM, Guo J, Newell JD, Jr., Hoffman EA, Sieren JC: Impact of advanced detector technology and iterative reconstruction on low-dose quantitative assessment of lung computed tomography density in a biological lung model. Medical physics 2018.
37. Dilger SKN: Pushing the boundaries: feature extraction from the lung improves pulmonary nodule classification. 2016.
38. Wheat LJ, Azar MM, Bahr NC, Spec A, Relich RF, Hage C: Histoplasmosis. Infectious disease clinics of North America 2016, 30(1):207-227.
39. Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B, Gareen IF, Gatsonis C, Goldin J, Gohagan JK et al: The National Lung Screening Trial: overview and study design. Radiology 2011, 258(1):243-253.
40. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD: Genetic epidemiology of COPD (COPDGene) study design. Copd 2010, 7(1):32-43.
41. Schwartz AG, Lusk CM, Wenzlaff AS, Watza D, Pandolfi S, Mantha L, Cote ML, Soubani AO, Walworth G, Wozniak A et al: Risk of Lung Cancer Associated with COPD Phenotype Based on Quantitative Image Analysis. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2016, 25(9):1341-1347.
42. Couper D, LaVange LM, Han M, Barr RG, Bleecker E, Hoffman EA, Kanner R, Kleerup E, Martinez FJ, Woodruff PG et al: Design of the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). Thorax 2014, 69(5):491-494.
43. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M et al: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging 2013, 26(6):1045-1057.
44. Armato SG, 3rd, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, Kazerooni EA, MacMahon H, van Beek EJ, Yankelevitz D et al: The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Academic radiology 2007, 14(11):1409-1421.
45. Armato SG, 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA et al: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics 2011, 38(2):915-931.

46. Athelogou M, Kim HJ, Dima A, Obuchowski N, Peskin A, Gavrielides MA, Petrick N, Saiprasad G, Colditz Colditz D, Beaumont H et al: Algorithm Variability in the Estimation of Lung Nodule Volume From Phantom CT Scans: Results of the QIBA 3A Public Challenge. Academic radiology 2016, 23(8):940-952.
47. Armato SG, 3rd, Drukker K, Li F, Hadjiiski L, Tourassi GD, Engelmann RM, Giger ML, Redmond G, Farahani K, Kirby JS et al: LUNGx Challenge for computerized lung nodule classification. Journal of medical imaging (Bellingham, Wash) 2016, 3(4):044506.
48. Armato SG, 3rd, Hadjiiski L, Tourassi GD, Drukker K, Giger ML, Li F, Redmond G, Farahani K, Kirby JS, Clarke LP: LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned. Journal of medical imaging (Bellingham, Wash) 2015, 2(2):020103.
49. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data 2017, 4:170177.
50. Uthoff J, Koehn N, Larson J, Dilger SK, Hammond E, Schwartz A, Mullan B, Sanchez R, Hoffman RM, Sieren JC et al: Post-Imaging Pulmonary Nodule Mathematical Prediction Models: Are They Clinically Relevant? . European radiology Accepted 2019.
51. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M et al: Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology 2017, 284(1):228-243.
52. Gould MK, Donington J, Lynch WR, Mazzone PJ, Midthun DE, Naidich DP, Wiener RS: Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013, 143(5 Suppl):e93S-e120S.
53. Mehta HJ, Mohammed TL, Jantz MA: The American College of Radiology Lung Imaging Reporting and Data System: Potential Drawbacks and Need for Revision. Chest 2017, 151(3):539-543.
54. Pinsky PF, Gierada DS, Black W, Munden R, Nath H, Aberle D, Kazerooni E: Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. Annals of internal medicine 2015, 162(7):485-491.
55. van Riel SJ, Ciompi F, Jacobs C, Winkler Wille MM, Scholten ET, Naqibullah M, Lam S, Prokop M, Schaefer-Prokop C, van Ginneken B: Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines. European radiology 2017, 27(10):4019-4029.
56. Gray EP, Teare MD, Stevens J, Archer R: Risk Prediction Models for Lung Cancer: A Systematic Review. Clinical lung cancer 2016, 17(2):95-106.

57. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES: The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Archives of internal medicine 1997, 157(8):849-855.
58. Gould MK, Ananth L, Barnett PG, Veterans Affairs SCSG: A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest 2007, 131(2):383-388.
59. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, Yasufuku K, Martel S, Laberge F, Gingras M et al: Probability of cancer in pulmonary nodules detected on first screening CT. The New England journal of medicine 2013, 369(10):910-919.
60. Li Y, Wang J: A mathematical model for predicting malignancy of solitary pulmonary nodules. World journal of surgery 2012, 36(4):830-835.
61. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, Hoekstra OS: Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. Chest 2005, 128(4):2490-2496.
62. Baldwin DR, Callister ME: The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. Thorax 2015, 70(8):794-798.
63. Hammer MM, Nachiappan AC, Barbosa EJM, Jr.: Limited Utility of Pulmonary Nodule Risk Calculators for Managing Large Nodules. Current problems in diagnostic radiology 2018, 47(1):23-27.
64. Al-Ameri A, Malhotra P, Thygesen H, Plant PK, Vaidyanathan S, Karthik S, Scarsbrook A, Callister ME: Risk of malignancy in pulmonary nodules: A validation study of four prediction models. Lung cancer (Amsterdam, Netherlands) 2015, 89(1):27-30.
65. Mehta HJ, Ravenel JG, Shaftman SR, Tanner NT, Paoletti L, Taylor KK, Tammemagi MC, Gomez M, Nietert PJ, Gould MK et al: The utility of nodule volume in the context of malignancy prediction for small pulmonary nodules. Chest 2014, 145(3):464-472.
66. Perandini S, Soardi GA, Motton M, Montemezzi S: Critique of Al-Ameri et al. (2015) - Risk of malignancy in pulmonary nodules: A validation study of four prediction models. Lung cancer (Amsterdam, Netherlands) 2015, 90(1):118-119.
67. Solitary Pulmonary Nodule Malignancy Risk (Mayo Clinic model) [https://reference.medscape.com/calculator/solitary-pulmonary-nodule-risk]
68. VA Clinical Model [https://magarray.com/calculator/]
69. Lung Cancer Risk Calculators [https://brocku.ca/lung-cancer-screening-and-risk-prediction/risk-calculators/]

70. Team RC: R: A language and environment for statistical computing. In. Edited by Computing RFfS. Vienna, Austria: R Core Team; 2019.
71. Chang W, Cheng J, Allaire J, Xie Y, McPherson J: shiny: Web Application Framework for R. In.; 2018.
72. Chung K, Mets OM, Gerke PK, Jacobs C, den Harder AM, Scholten ET, Prokop M, de Jong PA, van Ginneken B, Schaefer-Prokop CM: Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. Thorax 2018, 73(9):857-863.
73. McNitt-Gray MF, Kim GH, Zhao B, Schwartz LH, Clunie D, Cohen K, Petrick N, Fenimore C, Lu ZQ, Buckler AJ: Determining the Variability of Lesion Size Measurements from CT Patient Data Sets Acquired under "No Change" Conditions. Translational oncology 2015, 8(1):55-64.
74. Lin H, Huang C, Wang W, Luo J, Yang X, Liu Y: Measuring Interobserver Disagreement in Rating Diagnostic Characteristics of Pulmonary Nodule Using the Lung Imaging Database Consortium and Image Database Resource Initiative. Academic radiology 2017, 24(4):401-410.
75. Maiga AW, Deppen SA, Massion PP, Callaway-Lane C, Pinkerman R, Dittus RS, Lambright ES, Nesbitt JC, Grogan EL: Communication About the Probability of Cancer in Indeterminate Pulmonary Nodules. JAMA surgery 2018, 153(4):353-357.
76. Ferreira JR, Oliveira MC, de Azevedo-Marques PM: Characterization of Pulmonary Nodules Based on Features of Margin Sharpness and Texture. Journal of digital imaging 2017.
77. Zhao YR, van Ooijen PM, Dorrius MD, Heuvelmans M, de Bock GH, Vliegenthart R, Oudkerk M: Comparison of three software systems for semi-automatic volumetry of pulmonary nodules on baseline and follow-up CT examinations. Acta radiologica (Stockholm, Sweden : 1987) 2014, 55(6):691-698.
78. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Wang X, Yang H, Tan Y, Gillies R, Napel S: A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study. Journal of digital imaging 2016, 29(4):476-487.
79. Christe A, Bronnimann A, Vock P: Volumetric analysis of lung nodules in computed tomography (CT): comparison of two different segmentation algorithm softwares and two different reconstruction filters on automated volume calculation. Acta radiologica (Stockholm, Sweden : 1987) 2014, 55(1):54-61.
80. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P et al: Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. Tomography : a journal for imaging research 2016, 2(4):430-437.

81. Lampert TA, Stumpf A, Gancarski P: An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 2016.
82. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B et al: Fiji: an open-source platform for biological-image analysis. Nature methods 2012, 9(7):676-682.
83. MeVisLab: MeVisLab: A Development Environment for Medical Image Processing and Visualization. MeVis 2006.
84. Yushkevich PA, Gao Y, Gerig G: ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): 16-20 Aug. 2016 2016; 2016: 3342-3345.
85. Mukhopadhyay S: A Segmentation Framework of Pulmonary Nodules in Lung CT Images. Journal of digital imaging 2016, 29(1):86-103.
86. van Riel SJ: Malignancy risk estimation of pulmonary nodules in screening CTs: Comparison between a computer model and human observers. 2017, 12(11).
87. Sieren JP, Newell JD, Jr., Barr RG, Bleecker ER, Burnette N, Carretta EE, Couper D, Goldin J, Guo J, Han MK et al: SPIROMICS Protocol for Multicenter Quantitative Computed Tomography to Phenotype the Lungs. American journal of respiratory and critical care medicine 2016, 194(7):794-806.
88. Dhara AK, Mukhopadhyay S, Das Gupta R, Garg M, Khandelwal N: Erratum to: A Segmentation Framework of Pulmonary Nodules in Lung CT Images. Journal of digital imaging 2016, 29(1):148.
89. Galloway MM: Texture analysis using gray level run lengths. Computer graphics and image processing 1975, 4(2):172-179.
90. Chu A, Sehgal CM, Greenleaf JF: Use of gray value distribution of run lengths for texture analysis. Pattern Recognition Letters 1990, 11(6):415-419.
91. Dasarathy BV, Holder EB: Image characterizations based on joint gray level—run length distributions. Pattern Recognition Letters 1991, 12(8):497-502.
92. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J, Mari J-L: Shape and texture indexes application to cell nuclei classification. International Journal of Pattern Recognition and Artificial Intelligence 2013, 27(01):1357002.
93. Amadasun M, King R: Textural features corresponding to textural properties. IEEE Transactions on systems, man, and Cybernetics 1989, 19(5):1264-1274.

94. McCollough C, Bakalyar DM, Bostani M, Brady S, Boedeker K, Boone JM, Chen-Mayer HH, Christianson OI, Leng S, Li B et al: Use of Water Equivalent Diameter for Calculating Patient Size and Size-Specific Dose Estimates (SSDE) in CT: The Report of AAPM Task Group 220. AAPM report 2014, 2014:6-23.
95. Kaufman L, Rousseeuw PJ: Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis 1990:68-125.
96. Kaufman L, Rousseeuw P: Clustering by means of medoids: North-Holland; 1987.
97. Uthoff J, Sieren JC: Information Theory Optimization Based Feature Selection in Breast Mammography Lesion Classification. Conf Proc IEEE Eng Med Biol Soc 2018.
98. Vittinghoff E, McCulloch CE: Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. American Journal of Epidemiology 2007, 165(6):710-718.
99. Koay EJ, Ferrari M: Transport Oncophysics in silico, in vitro, and in vivo. Preface. Physical biology 2014, 11(6):060201.
100. Hammond E, Sloan C, Newell JD, Jr., Sieren JP, Saylor M, Vidal C, Hogue S, De Stefano F, Sieren A, Hoffman EA et al: Comparison of low- and ultralow-dose computed tomography protocols for quantitative lung and airway assessment. Medical physics 2017, 44(9):4747-4757.
101. Azar MM, Hage CA: Clinical Perspectives in the Diagnosis and Management of Histoplasmosis. Clinics in chest medicine 2017, 38(3):403-415.
102. Pinsky PF, Gierada DS, Nath PH, Kazerooni E, Amorosa J: National lung screening trial: variability in nodule detection rates in chest CT studies. Radiology 2013, 268(3):865-873.
103. Warren WA, Markert RJ, Stewart ED: Pulmonary nodule tracking using chest computed tomography in a histoplasmosis endemic area. Clinical imaging 2015, 39(3):417-420.
104. Manos NE, Ferebee SH, Kerschbaum WF: Geographic variation in the prevalence of histoplasmin sensitivity. Diseases of the chest 1956, 29(6):649-668.
105. Cicchetti DV: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994, 6(4):284-290.
106. Cohen J: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin 1968, 70(4):213-220.
107. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG: Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. Translational oncology 2015, 8(6):524-534.
108. Al-Kadi OS: Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed tomography lung tumour images. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society 2010, 34(6):494-503.

109. Wasswa-Kintu S, Gan WQ, Man SF, Pare PD, Sin DD: Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. Thorax 2005, 60(7):570-575.
110. Wilson DO, Leader JK, Fuhrman CR, Reilly JJ, Sciurba FC, Weissfeld JL: Quantitative computed tomography analysis, airflow obstruction, and lung cancer in the pittsburgh lung screening study. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2011, 6(7):1200-1205.
111. Maldonado F, Bartholmai BJ, Swensen SJ, Midthun DE, Decker PA, Jett JR: Are airflow obstruction and radiographic evidence of emphysema risk factors for lung cancer? A nested case-control study using quantitative emphysema analysis. Chest 2010, 138(6):1295-1302.
112. Chubachi S, Takahashi S, Tsutsumi A, Kameyama N, Sasaki M, Naoki K, Soejima K, Nakamura H, Asano K, Betsuyaku T: Radiologic features of precancerous areas of the lungs in chronic obstructive pulmonary disease. International journal of chronic obstructive pulmonary disease 2017, 12:1613-1624.
113. Aamli Gagnat A, Gjerdevik M, Gallefoss F, Coxson HO, Gulsvik A, Bakke P: Incidence of non-pulmonary cancer and lung cancer by amount of emphysema and airway wall thickness: a community-based cohort. The European respiratory journal 2017, 49(5).
114. Gierada DS, Guniganti P, Newman BJ, Dransfield MT, Kvale PA, Lynch DA, Pilgram TK: Quantitative CT assessment of emphysema and airways in relation to lung cancer risk. Radiology 2011, 261(3):950-959.
115. Wille MM, Thomsen LH, Petersen J, de Bruijne M, Dirksen A, Pedersen JH, Shaker SB: Visual assessment of early emphysema and interstitial abnormalities on CT is useful in lung cancer risk analysis. European radiology 2016, 26(2):487-494.
116. Bae K, Jeon KN, Lee SJ, Kim HC, Ha JY, Park SE, Baek HJ, Choi BH, Cho SB, Moon JI: Severity of pulmonary emphysema and lung cancer: analysis using quantitative lobar emphysema scoring. Medicine 2016, 95(48):e5494.
117. Johannessen A, Skorge TD, Bottai M, Grydeland TB, Nilsen RM, Coxson H, Dirksen A, Omenaas E, Gulsvik A, Bakke P: Mortality by level of emphysema and airway wall thickness. American journal of respiratory and critical care medicine 2013, 187(6):602-608.
118. Monticciolo DL, Newell MS, Hendrick RE, Helvie MA, Moy L, Monsees B, Kopans DB, Eby PR, Sickles EA: Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging. Journal of the American College of Radiology : JACR 2017, 14(9):1137-1143.

119. Sedgwick EL, Ebuoma L, Hamame A, Phalak K, Ruiz-Flores L, Ortiz-Perez T, Sepulveda KA: BI-RADS update for breast cancer caregivers. Breast cancer research and treatment 2015, 150(2):243-254.
120. Hapfelmeier A, Horsch A: Image feature evaluation in two new mammography CAD prototypes. International journal of computer assisted radiology and surgery 2011, 6(6):721-735.
121. Jiao Z, Gao X, Wang Y, Li J: A parasitic metric learning net for breast mass classification based on mammography. Pattern Recognition 2017.
122. Li H, Meng X, Wang T, Tang Y, Yin Y: Breast masses in mammography classification with local contour features. Biomedical engineering online 2017, 16(1):44.
123. Xie W, Li Y, Ma Y: Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing 2016, 173:930-941.
124. Zhang Y, Tomuro N, Furst J, Stan Raicu D: Using BI-RADS Descriptors and Ensemble Learning for Classifying Masses in Mammograms. In: Medical Content-Based Retrieval for Clinical Decision Support: First MICCAI International Workshop, MCBR-CDS 2009, London, UK, September 20, 2009, Revised Selected Papers. edn. Edited by Caputo B, Müller H, Syeda-Mahmood T, Duncan JS, Wang F, Kalpathy-Cramer J. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010: 69-76.
125. Zheng Y, Keller BM, Ray S, Wang Y, Conant EF, Gee JC, Kontos D: Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Medical physics 2015, 42(7):4149-4160.
126. Sun W, Tseng TL, Qian W, Zhang J, Saltzstein EC, Zheng B, Lure F, Yu H, Zhou S: Using multiscale texture and density features for near-term breast cancer risk analysis. Medical physics 2015, 42(6):2853-2862.
127. Li H, Mendel KR, Lan L, Sheth D, Giger ML: Digital Mammography in Breast Cancer: Additive Value of Radiomics of Breast Parenchyma. Radiology 2019:181113.
128. Tan M, Pu J, Zheng B: Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. International journal of computer assisted radiology and surgery 2014, 9(6):1005-1020.
129. Qiu Y, Yan S, Gundreddy RR, Wang Y, Cheng S, Liu H, Zheng B: A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. Journal of X-ray science and technology 2017, 25(5):751-763.
130. Sawyer Lee R, Gimenez F, Hoogi A, Rubin D: Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive. In.; 2016.

131. Chan TF, Vese LA: Active contours without edges. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 2001, 10(2):266-277.
132. Ganesan K, Acharya UR, Chua CK, Min LC, Abraham KT, Ng KH: Computer-aided breast cancer detection using mammograms: a review. IEEE reviews in biomedical engineering 2013, 6:77-98.
133. Jalalian A, Mashohor S, Mahmud R, Karasfi B, Saripan MIB, Ramli ARB: Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. EXCLI journal 2017, 16:113-137.
134. Yassin NIR, Omran S, El Houby EMF, Allam H: Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Computer methods and programs in biomedicine 2018, 156:25-45.
135. Abbas Q: DeepCAD: A computer-aided diagnosis system for mammographic masses using deep invariant features. Computers 2016, 5(4):28.
136. Verma B, McLeod P, Klevansky A: A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. Pattern Recognition 2009, 42(9):1845-1852.
137. Jaffar MA: Deep learning based computer aided diagnosis system for breast mammograms. International Journal of Advanced Computer Science and Applications 2017, 8(7):286-290.
138. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B, Sanchez CI: A survey on deep learning in medical image analysis. Medical image analysis 2017, 42:60-88.
139. Gerard SE, Patton TJ, Christensen GE, Bayouth JE, Reinhardt JM: FissureNet: A Deep Learning Approach For Pulmonary Fissure Detection in CT Images. IEEE transactions on medical imaging 2019, 38(1):156-166.
140. Jiang H, Ma H, Qian W, Gao M, Li Y, Hongyang J, He M, Wei Q, Mengdi G, Yan L: An Automatic Detection System of Lung Nodule Based on Multigroup Patch-Based Deep Learning Network. IEEE journal of biomedical and health informatics 2018, 22(4):1227-1237.
141. Davis J, Goadrich M: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, USA: ACM; 2006: 233-240.
142. Fan J, Upadhye S, Worster A: Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine 2006, 8(1):19-20.
143. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 2011, 12(1):77.

144. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988, 44(3):837-845.
145. Demler OV, Pencina MJ, D'Agostino RB, Sr.: Misuse of DeLong test to compare AUCs for nested models. Statistics in medicine 2012, 31(23):2577-2587.
146. Youden WJ: Index for rating diagnostic tests. Cancer 1950, 3(1):32-35.
147. McNemar Q: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947, 12(2):153-157.
148. Jarque CM, Bera AK: Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics letters 1980, 6(3):255-259.
149. Wilcoxon F: Some rapid approximate statistical procedures. Annals of the New York Academy of Sciences 1950, 52(1):808-814.
150. Fisher RA: On the Interpretation of χ² from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 1922, 85(1):87-94.
151. Pearson K: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1900, 50(302):157-175.
152. Kymes SM, Lee K, Fletcher JW: Assessing diagnostic accuracy and the clinical value of positron emission tomography imaging in patients with solitary pulmonary nodules (SNAP). Clinical Trials 2006, 3(1):31-42.
153. Boykov Y, Veksler O, Zabih R: Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence 2001, 23(11):1222-1239.
154. Lassen BC, Jacobs C, Kuhnigk JM, van Ginneken B, van Rikxoort EM: Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans. Physics in medicine and biology 2015, 60(3):1307-1323.
155. Shen S, Bui AA, Cong J, Hsu W: An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy. Computers in biology and medicine 2015, 57:139-149.
156. Huttenlocher DP, Klanderman GA, Rucklidge WJ: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 1993, 15(9):850-863.
157. Chatfield C, Collins AJ: Principal component analysis. In: Introduction to multivariate analysis. edn.: Springer; 1980: 57-81.

158. Peduzzi P, Concato J, Feinstein AR, Holford TR: Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of clinical epidemiology 1995, 48(12):1503-1510.
159. McGill W: Multivariate information transmission. Transactions of the IRE Professional Group on Information Theory 1954, 4(4):93-111.
160. Ho TK: Random decision forests. In: Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on: 1995: IEEE; 1995: 278-282.
161. Sonka M, Hlavac V, Boyle R: Image processing, analysis, and machine vision: Cengage Learning; 2014.
162. Wolpert DH: Stacked generalization. Neural Networks 1992, 5(2):241-259.
163. Dietterich TG: Ensemble methods in machine learning. In: International workshop on multiple classifier systems: 2000: Springer; 2000: 1-15.

APPENDIX A: STATISTICAL METHODS

This section details the performance and statistical comparison techniques used in this dissertation. The reported measures provide assessment of the results, additional performance and statistical measures exist but were not applied here for the sake of clarity and consistency.

A.1. Classifier Performance Measures

Tool performance was assessed using area-under the receiver operating characteristic curve (AUC-ROC) and area-under the precision-recall curve (AUC-PR)[141] .

A.1.1. Receiver-Operator Characteristic Curve

Area-under-the-curve of the receiver-operator characteristic (AUC-ROC), also known as the c-statistic, is a common binary classification assessment measure. It is equal to the probability that a higher risk will be assigned to a randomly chosen true high-risk case (here, cancer) than a randomly chosen low risk case (here, non-cancer)[141-143]. The curve itself plots the tradeoff between the true positive rate against the false positive rate for every possible threshold of predicted risk (between 0 and 1); AUC-ROC is calculated as the integration of the plotted curve. Extreme values of AUC-ROC indicate: 1 as complete true separation of classes (all true high-risk cases predicted at greater risk than all true low-risk cases), 0 as complete false separation of classes (all true high-risk cases predicted at lower risk than all true low-risk cases, note: still perfect classification can flip prediction about risk = 0.5 for AUC-ROC = 1), and 0.5 as random chance (prediction no better than randomly assigning risk).

A.1.1.1. Delong – Comparison of ROC curves

The Delong approach is a nonparametric method of evaluating and comparing the performance of diagnostic tests using theory on generalized U-statistics[144]. The 95% confidence interval for AUC-ROC can be assessed using bootstrapping methods of the Delong approach. The ROC of two different classifiers on the same subjects can be compared for statistical difference. A caveat to this method has been raised in Delmer et al. with regards to this comparison on 'nested' risk prediction models – essentially, two models: one built with a full set of N features and the other built with a subset of the N features[145]. In this work, the Delong approach is only used to compare models built with sufficiently different predictor variables.

A.1.2. Precision-Recall Curve

AUC-ROC assessments can be influenced by class imbalance (i.e. a dataset with different proportions of high-risk vs low-risk classes). The precision-recall curve and associated integrated area-under (AUC-PR) can provide a more robust analysis of performance. Similar to AUC-ROC, AUC-PR is

taken as the integration of the plot of precision against recall. Extreme values of AUC-PR indicate: 1 as perfect classifier with complete true separation of the classes and prediction values at the binary poles and 0 as a poor classifier.

A.1.3. Categorizing Risks - Thresholds

While AUC-ROC and AUC-PR are robust measures of classifier performance, eventually for clinical utility, a set threshold must be selected as a 'cut-off' point which will binarize the predicted risk into discrete categories. Once a threshold has been selected, measures of accuracy, sensitivity, and specificity can be calculated.

A.1.3.1. Rounding Prediction Threshold

As binary classifier prediction results are trained to provide a floating-point number between 0 and 1, rounding to the nearest whole number is a natural thresholding method. Here, all cases with a prediction level below 0.50 are classified as benign and all above 0.50 are classified as malignant. This is the method that was used in the prior approach and was carried through in the feature reduction and selection methodology development sections (**Appendices E and F)**.

A.1.3.2. Youden Threshold

The Youden threshold is based on the optimal value of the Youden J Statistic[146]. The Youden J statistic finds the optimal point for the sensitivity-specificity tradeoff:

$$J = sensitivity + specificity - 1$$

[Equation A.1.3.1]

In a perfect classification (AUC-ROC = 1, AUC-PR = 1), the Youden has near-infinite possibilities (as all conceivable thresholds between >0 and < 1), in this case the Youden can be set a 0.5.

A.1.3.3. Custom-Calibrated Thresholds

The selection of a threshold can be heuristically done to suit the needs and calibration requirements of the model. This can be done by simply selecting the threshold that meets the desired requirements in the training cohort. For example, the user may wish to apply a threshold aiming to achieve 90% sensitivity. To do this, the AUC-ROC curve can be interrogated for a threshold that achieves 90% sensitivity in the training data can be determined.

A.1.3.4. McNemar's – Classification Error of Threshold or Categorized Risk

McNemar's test is a statistical test used on paired nominal data to determine whether there is marginal homogeneity in classification score between two tools[147]. The test is applied to a 2x2 contingency table as follows:

McNemar's Contingency Table		Tool 1	
		Correct	Incorrect
Tool 2	Correct	A	B
	Incorrect	C	D

The McNemar test statistic is:

$$\chi^2 = \frac{(B - C)^2}{B + C}$$

[Equation A.1.3.3.1]

This was modified for continuity correction by Edwards to approximate the binomial exact p-value:

$$\chi^2 = \frac{(|B - C| - 1)^2}{B + C}$$

[Equation A.1.3.3.2]

X^2 has a chi-squared distribution with 1 degree of freedom.

A.1.4. Threshold-based Performance Measures

From a threshold, several performance measures can be assessed including:

Accuracy:

$$accruacy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

[Equation A.1.4.1]

Sensitivity (also known as: recall, hit rate, or true positive rate):

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative}$$

[Equation A.1.4.2]

Specificity (also known as: selectivity or true negative rate):

$$specificity = \frac{true\ negative}{true\ negative + false\ positive}$$

[Equation A.1.4.3]

A.2: Variable/Features Statistic Differences

A.2.1. Continuous Variables

For continuous variables, normality is assessed using the Jarque-Bera goodness-of-fit test[148]. For variables that conform to normality, statistical difference of is assessed with either a two-sample t-test.

Paired t-tests operate under the null hypothesis that the means are equal; non-paired t-tests operate under the null hypothesis that the mean differences are equal.
For variables that fail the Jarque-Bera test, either the Wilcoxon signed rank test (paired) or Wilcoxon rank sum test (non-paired) is used to assess statistical differences in groups[149]. The Wilcoxon signed rank test operates under the null hypothesis that the medians are equal; the Wilcoxon rank sum test operates under the null hypothesis that the median differences are equal.

A.2.2. Discrete or Categorical Variables

A.2.2.1. Fisher's Exact Test

For nominal variables, Fisher's exact test is a method for determining if there are statistical non-random associations from contingency tables[150]. Fisher's exact test operates under the null hypothesis that the proportions are equal.

A.2.2.2. Persons' chi-squared test

The Person's chi-squared test is a statistical test for unpaired nominal data used to assess goodness-of-fit, homogeneity, and independence which uses contingency tables[151]. It is suitable for use when at least 75% if cells of the contingency table are expected to have counts greater than five. Person's chi-squared test operates under the null hypothesis that the proportions are equal.

A.3. Data Partitioning Methods

A.3.1. Training/Validation

The ultimate goal of a classification tool is to be employed on cases were the true classification is unknown. As such, it is imperative to examine how a tool performs on new, non-development cases. A common method to achieve this performance analysis is splitting the dataset into a training set and a validation set. The classification tool is developed using only the training set and the finalized tool is tested on the validation set; in this schema, the tool has been blinded to the true classification of the validation subjects allowing for quasi-real-world performance measures to be obtained.
This process has limitations. As the training dataset is only a subset of the available data, the development of the tool is potentially only seeing a fraction of the true dataset variance. This is a particular issue in datasets that are small or that contain a large variability in subject presentation; in these cases, the developed tool is more likely to either be over-trained on the development dataset or to perform poorly on new cases that have variability not seen in the development dataset.

A.3.2. K-fold Cross Validation

K-fold Cross validation (kCV) is a dataset partitioning method that seeks to alleviate the limitations of the training/validation method and utilize the entire dataset both as training and as validation cases. kCV is performed by randomly portioning the dataset into k-groups of subjects. Then, the classifier development and validation steps are repeated k-times with each time a different group left-out of development and used for validation. In this method, the performance measures (AUC-ROC, AUC-PR, sensitivity, specificity, etc.) are calculated from the group's validation-run predictions. The selection of k is often limited by computational load of classifier development. Small values of k reduce the number of times the development-validation steps need to be run, but it increases the limitations seen in **A.3.1** as the classifier is being trained on fewer cases. Large values of k increase the number of cases used in training a classifier but require more runs of the development-validation steps which can be prohibitive in classification development pipelines that are computationally complex. The most common k used in kCV is 10.

A.3.3. Leave-one-out (Extreme k-fold Cross Validation)

Leave-one-out (LOO) is an extreme form of kCV where k is equal N, to the number of subjects. The development-validation pipeline is run N-times. In this method, the performance measures (AUC-ROC, AUC-PR, sensitivity, specificity, etc.) are calculated from the subjects left-out validation-run predictions. This method utilizes the most data for development (N-1 cases) which makes it efficient in dataset variability inclusion; however, in cases were classifier development takes a significant amount of time and/or computational power it can be prohibitive. LOO is most suited to small datasets.

APPENDIX B: MATHEMATICAL PREDICTION MODELS

B.1. Model Formulas

For each model **x** is calculated such that the pre-test probability of malignancy is $\frac{e^{x}}{(1+e^{x})}$

B.1.1. Mayo Clinic (MC) Model

The Mayo Clinic (MC) MPM, published 1997, was developed on 629 subjects with nodules (146 malignant, 406 benign, 77 indeterminant) using clinical radiograph and CT scans[57]. It is specified using the following definition of **x**:

$$x = -6.8272 + 0.0391 * \boldsymbol{Age} + 0.7917 * \boldsymbol{Smoker} + 1.3388 * \boldsymbol{CancerHx} + 0.1274 * \boldsymbol{Diameter} + 1.0407 * \boldsymbol{Spiculation} + 0.7838 * \boldsymbol{UpperLobe}$$

[Equation B.1.1]

In this equation, ***Age*** is the patient's age in years, ***Smoker*** (binary) equals 1 if the patient has a history as a current or former smoker (otherwise = 0), ***CancerHx*** (binary) is 1 if the patient has a history of extrathoracic cancer that was diagnosed more than 5 years ago (otherwise = 0), ***Diameter*** is the maximum in-plane diameter of the nodule in millimeters, ***Spiculation*** (binary) is 1 if the edge of the nodule is spiculated (otherwise = 0), and ***UpperLobe*** (binary) is 1 if the nodule is located in the left upper lobe (LUL) or right upper lobe (RUL) (otherwise = 0).

B.1.2. United States Department of Veterans Affairs (VA) Model

The U.S. Department of Veterans Affairs (VA) MPM, published 2007, was developed on 375 subjects with nodules (204 malignant, 171 benign) using CT scans acquired as part of a prospective study comparing the effectiveness of PET-CT and CT[58,152]. It is specified using the following definition of **x**:

$$x = -8.404 + \left(0.779 * \frac{\boldsymbol{Age}}{10}\right) + (2.061 * \boldsymbol{Smoker}) + (0.112 * \boldsymbol{Diameter}) - (0.567 * \boldsymbol{CessationTime})$$

[Equation B.1.2]

In this model, ***Age*** is the patient's age in years, ***Smoker*** (binary) equals 1 if the patient has a history as a current or former smoker (otherwise = 0), ***Diameter*** is the diameter of the nodule in millimeters, and ***CessationTime*** is the number of years since quitting smoking.

B.1.3. Peking University (PU) Model

The Peking University (PU) MPM, published 2012, was developed on 375 subjects with nodules (229 malignant, 142 benign) using clinical radiograph and CT scans[60]. It is specified using the following definition of **x**:

$$x = -4.496 + (0.07 * \boldsymbol{Age}) + (0.676 * \boldsymbol{Diameter}/10) + (0.736 * \boldsymbol{Spiculation}) + (1.267 * \boldsymbol{Family\ History\ of\ Cancer}) - (1.615 * \boldsymbol{Calcification}) - (1.408 * \boldsymbol{Border})$$

[Equation B.1.3]

Here, ***Age*** is the patient's age in years, ***Diameter*** is the diameter of the nodule in millimeters, and ***Spiculation***, ***Family History of Cancer***, ***Calcification***, and ***Border*** (all binary) are 1 if the risk variable is present (otherwise = 0).

B.1.4. Brock University (BU) Model

The Brock University (Brock) MPM, published 2013, was developed on 1871 subjects with nodules (102 malignant, 1769 benign) using low-dose screening CT scans[59]. It is specified using the following definition of **x**:

$$x = \left(0.0287 * (\boldsymbol{Age} - 62)\right) + (0.6011 * \boldsymbol{Sex}) + (0.2961 * \boldsymbol{Family\ History\ of\ Lung\ Cancer}) + (0.2953 * \boldsymbol{Emphysema}) - \left[5.3854 * \left(\frac{\boldsymbol{Nodule\ Size}}{10}\right)^{-0.5} - 1.58113883\right] + \boldsymbol{Nodule\ Type} + (0.6581 * \boldsymbol{UpperLobe}) - \left(0.0824 * (\boldsymbol{Nodule\ Count} - 4)\right) + (0.7729 * \boldsymbol{Spiculation}) - 6.7892$$

[Equation B.1.4]

Here, ***Age*** is the patient's age in years, ***Sex*** is 1 if the patient is female or 0 if the patient is male, ***Family History of Lung Cancer*** is 1 if there is a family history of lung cancer (otherwise = 0), ***Emphysema*** is 1 if emphysematous changes is noted in the lungs (otherwise = 0), ***Nodule Size*** is the diameter of the nodule in millimeters, ***Nodule Type*** is -0.1276 if the nodule is nonsolid or ground-glass, 0.377 if the nodule is partially solid, or 0 if the nodule is solid, ***UpperLobe*** is 1 if nodule is located in the left upper lobe (LUL) or right upper lobe (RUL) (otherwise = 0), ***Nodule count*** is number of nodules noted in the lungs, and ***Spiculation*** is 1 if the nodule is spiculated (otherwise = 0).

B.2. Youden Threshold Stability and Calibration Set Size Algorithm

The following section describes the methods and results of determining an adequate calibration cohort size for Youden threshold stability.

B.2.1. Median Absolute Deviation

Median absolute deviation (MAD) was selected as the measure to determine adequate calibration cohort size as it is relatively insensitive to outliers and considers the full unsigned deviation from the set median.

$$median\big(\, abs(\, trial_{threshold} - median_{set\ threshold})\big)$$

[Equation B.2.1]

Here, the $\boldsymbol{trial_{threshold}}$ is the threshold calculated for a single trial and the $\boldsymbol{median_{set\ threshold}}$ is the median threshold across all trials with a given cohort size.

B.2.2. Algorithm for determining Youden threshold stability

Algorithm (Pseudo-code) for determining the Youden threshold stability:

```
For each N = 50 by 5 to 250
        For each trial in 10,000
                Trial_sample = Random sample N subjects from the full cohort without replacement
                Trial_Youden = Calculate Youden threshold for trial sample
        End
N_median = median of Trial_Youden
        N_MAD = median of the absolute of Trial_Youden – N_median
End
Find arg-min(N_MAD < 0.05)
```

APPENDIX C: SEMI-AUTOMATIC SEGMENTATION METHODS

C.1. Semi-automated Segmentation Pipelines

C.1.1. FIJI-ImageJ (FIJI)

The FIJI segmentation was performed by processing through a series of built-in FIJI plugins (free download available from: https://fiji.sc/). The CT digital imaging and communications in medicine (DICOM) dataset was imported into FIJI and the windowing level for the image was adjusted for improved lung tissue contrast (level = -600 HU, window = 1600). Using the 3D Crop Plugin, the VOI was selected to match the manual segmentation. For parenchyma segmentation the image was processed to a binary mask using the Otsu Threshold Method (18). For nodule segmentation, the VOI was first inspected for obstructions interacting with the nodule; if an attachment occurred then the offending voxels were removed using the Drawing Tool to set their value to background. Once the nodule was isolated, 3D Manual Spot segmentation was done by placing a seed in the center of the nodule and applying a Classical Gauss fitting threshold whose parameters, radius (1-3.5) and standard deviation (0.5-3.0), were manually adjusted to contain only the nodule voxels.

C.1.2. MeVisLab (MVL)

The MVL segmentation was performed by building a pipeline consisting of existing MVL plugins (free download available from: http://www.mevislab.de/mevislab/). The CT DICOM dataset was imported into the MVL platform and each VOI selected to correspond with the manual segmentation. For parenchymal segmentation, a seed was added to the nodule and all non-parenchymal lung objects (chest wall, vessels). These regions were grown using a 3D 6-Neighborhood relation with smart region growing at an interval size of 5%. A binary mask of the region was generated and saved as the valid parenchyma mask. For nodule segmentation, the previous region-grown mask was analyzed for nodule obstructions and offending voxels were removed by setting their value to background. Using this mask, a seed was added to the center of the nodule and again underwent region-growing using an interval size of 5%.

C.1.3. ITK-Snap (ITK-S)

The segmentation in the ITK-S environment (http://www.itksnap.org/pmwiki/pmwiki.php) did not require any pipeline development outside of the Select Active Contour Segmentation (SNAKE) function. The CT DICOM dataset was imported into ITK-S, and the VOI was selected to correspond with the manual segmentation VOI. For parenchyma segmentation preprocessing was performed using the clustering mode of SNAKE, and bubbles of radius 10 were placed throughout the parenchyma including at least one for every three slices in the VOI. The active contour was evolved until the parenchyma was encompassed. For nodule segmentation, SNAKE was also used; the dataset was pre-processed by

adjusting the lower threshold to ~-500 HU and setting the upper threshold to the HU maximum. The segmentation was initialized by placing a bubble in the centroid of the nodule and adjusting the radius of the bubble so as to not exceed the nodule boundary. The active contour was evolved until the nodule was encompassed. If the irregular shape of the nodule provided a poor fit using the default settings of SNAKE, then adjustments were made to the region competition and smoothing forces which act as cost functions to emphasize connectivity and smooth geometries.

C.1.4. Mukhopadhyay-MatLab (ML)

This method followed the segmentation framework for solid pulmonary nodules described by Mukhopadhyay[85,88]. It consisted of a MatLab script containing a pipeline of built in functions. The CT DICOM data was imported to MatLab and a VOI selected to correspond with the manual segmentation. A seed point was selected at the centroid of the nodule. To generate the nodule mask, the VOI was pre-processed by thresholding at -500 HU followed by connected components analysis to generate a foreground image of the nodule and any attachments. In this method, pleural attachments were removed through identification of points on the nodule boundary by ray casting from the seed point and convex hull operation on a fitted ellipsoid enclosing the nodule, and vessel attachments were removed through pruning of geodesic distance maps. To generate the parenchyma mask, the VOI was processed by intensity thresholding up to -500 HU.

C.1.5. Graph-cuts (GC)

Building upon the strengths of the other assessed pipelines, an in-house method was developed in MatLab, incorporating graph cuts. A seed point was placed at the nodule centroid and several (1-10) were placed in the lung parenchyma. Using a closed 26-neighborhood connected component analysis and Otsu's thresholding, the lungs were roughly identified. The maximum length of lungs from posterior to anterior was found and the midpoint of the lung was identified. From the midpoint, 15% of the maximum length was dilated. The masked image was smoothed using a curvature anisotropic diffusion filter. A graph was constructed using image intensity and nodule size considerations: the boundary term coming from the local image gradient and the regional term incorporating two properties of the nodule, (1) intensity-based probability (assumed means of the background and object based on prior knowledge) and (2) a distance-based probability based on geodesic distance from the seed point. This graph was then run through a graph-cuts algorithm employing the fast-continuous max-flow method proposed by J. Yuan and Y. Boykov[153]. Post-processing of nodule attachments was done using the approaches of Mukhopadhyay explained in the ML section above.

C.2. Perinodular Parenchyma Rings and Bands Segmentation Methods

As we hypothesize nodules to have a size-proportional effect on their surrounding structures, parenchymal rings were taken as a percentage of the nodule maximum in-plane diameter. We investigated exclusive parenchyma bands at each of the quartiles (25%-band, 50%-band, 75%-band, and 100%-band) and inclusive parenchyma rings at each of the quartiles (25%-ring, 50%-ring, 75%-ring, and 100%-ring); **Figure C.1** shows the production of the parenchyma bands compared to rings. Parenchymal features pulled from the bands will be compared between band sizes and to the previous results using inclusive parenchymal rings. This seeks to determine if the significant signal is coming from the exclusive or inclusive parenchymal percent.

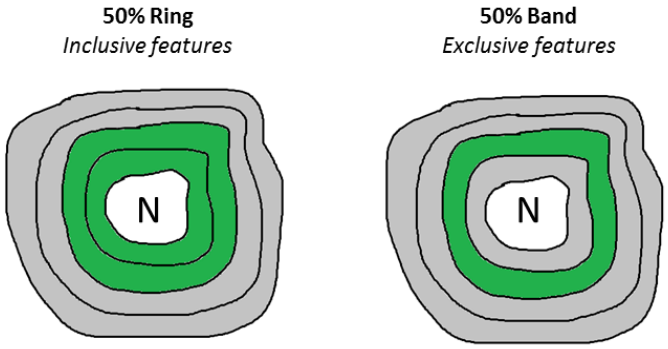

Figure C.1: Pictorial representation of parenchymal rings and parenchymal bands. Green indicates the region of feature extraction, N represents the corresponding pulmonary nodule.

C.3. Segmentation Performance Analysis

Tool performance was analyzed using five measures: sensitivity, specificity, Jaccard distance, volumetric error rate, and scaled Hausdroff distance. Let A be the nodule truth mask, B be the nodule mask resulting from a semi-automated tool. Similarly, let C be the background truth mask and D be the background mask resulting from a semi-automated tool.

C.3.1. Sensitivity and Specificity

The sensitivity and specificity have been used to assess segmentation accuracy compared to a truth with 1 being completely accurate and 0 being no accuracy. In sensitivity and specificity, the overlap of correctly labeled voxels is summarized.

$$Sensitivity = \frac{|A \cap B|}{|A \cap B| + |C \cap B|}$$

[Equation C.2.1.1]

$$Specificity = \frac{|C \cap D|}{|C \cap D| + |A \cap D|}$$

[Equation C.2.1.2]

C.3.2. Jaccard Distance

The Jaccard distance (JD) has been used to measure the overlap similarity between segmentations in several other publications with 1 being no overlap and 0 being complete overlap[154]. The JD uses the presence of the data while ignoring information about the data abundance. It is calculated by subtracting from one the ratio of the size of the intersection and the size of the union of the two masks (A - truth and B - tool):

$$JD_{nodule}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

[Equation C.2.2]

C.3.3. Volumetric Error Rate

The volumetric error rates (VE) have been used to indicate base-level variation in segmentation sizes. The result ranges from -100% to 100% with 0% error being optimal [155]. Unlike the JD, the VE indicates the direction of the error, with negative errors showing an underestimation by the tool and positive errors an overestimation. The VE is not a direct metric of mask overlap or spatial similarity but rather a metric of size similarity.

$$VE_{segmentation} = \frac{V_{tool} - V_{truth}}{V_{truth}}\%$$

[Equation C.2.3]

C.3.4. Scaled Hausdorff Distance

The Hausdorff distance (HD) is a measure of the differential distance between corresponding edge points on two masks[156].

$$H(A,B) = \underset{a \in A}{max} \left\{ \underset{b \in B}{min} \{ d(a,b) \} \right\}$$

[Equation C.2.4.1]

$$SHD = H(A,B)/\max(H(A,B), cohort)$$

[Equation C.2.4.2]

Here, ***d(a,b)*** is the distance from pixel *a* in mask A to pixel *b* in mask B. The minimum distance between *a* and *b* seen over all the edge pixels *b* is calculated for each pixel *a*; this corresponds to the corresponding closest edge pixel on the scans. The maximum of these minimum distances over all the edge pixels is the maximum HD (MHD). To that end, an MHD value of 0 equates to two identical masks and a HD value approaching infinity equates to dissimilar masks. The major advantage of MHD is that because it demonstrates the differences in only the edges of the segmentation, it is not influenced as much by the bulk of the mask (interior voxels), which constitute a large sway in the JD and VE. For comparison among tools, Hausdroff Distance was scaled (SHD) to be from 0 to 1 as in **Equation C.2.4.2**, with 1 being the maximum MHD seen in the complete cohort of segmentations.

APPENDIX D: FEATURE EXTRACTION

The feature set was expanded to include additional imaging features, specifically texture and new size measures novel to classification schema.

D.1. Intensity Features

Five intensity histogram (IH) features were added. The image voxel value at percentiles 5^{th}, 25^{th}, 7^{th}, and 95^{th}. The proportion of voxels about 100 HU was included to potentially identify nodules with calcifications[80].

- Percentile 5^{th}, 25^{th}, 75^{th}, 95^{th}
- Proportion Above 100 HU

D.2. Gray-Level Run-Length Textures

Gray-level run length (GLRL) features have been widely used in the arena of medical imaging feature extraction[89-91]. A run is a set of consecutive, collinear voxels with the same gray level value. Runs are calculated in all of the principle directions (0°,45°, 90° ,135°). From these matrices' measures of coarse texture (long runs) and fine texture (short runs) can be extracted as well as gray-level and run non-uniformity which could describe texture heterogeneity.

- Short Run Emphasis
- Long Run Emphasis
- Gray-Level Non-uniformity
- Run-Length Non-uniformity
- Run Length Percentage
- Low Gray-Level Run Emphasis
- High Gray-Level Run Emphasis
- Short Run Low Gray-Level Emphasis
- Short Run High Gray-Level Emphasis
- Long Run Low Gray-Level Emphasis
- Long Run High Gray-Level Emphasis
- Gray-Level Variance
- Run-Length Variance

D.3. Gray-Level Size-Zone Textures

There are 13 gray-level size zone texture features (GLSZ) were generated from size zone matrix built under run length matrix principles where the value of the matrix at a (row, column) is equal to the (gray levels, number of zones of a size)[92]. This results in a matrix size of number of gray levels by a quantization of the size of the largest zone; a heterogeneous texture would result in a tall and thin matrix while a homogenous texture matrix would be short and wide. These additional features include measures of emphasis, non-uniformity, and variance in the size and distribution of the gray-level size zone matrix.

- Small Zone Emphasis
- Large Zone Emphasis
- Gray-Level Non-uniformity
- Zone-Size Non-uniformity
- Zone Percentage
- Low Gray-Level Zone Emphasis
- High Gray-Level Zone Emphasis
- Small Zone Low Gray-Level Emphasis
- Small Zone High Gray-Level Emphasis
- Large Zone Low Gray-Level Emphasis
- Large Zone High Gray-Level Emphasis
- Gray-Level Variance
- Zone-Size Variance

D.4. Neighborhood Gray-Tone Difference Matrix Textures

Neighborhood Gray-Tone Difference Matrix Textures (NGTD) are a set of five features whose derivations were heuristically developed to assess for texture types within a region[93]. Five texture features calculated from the neighborhood gray-tone difference (NGTD) matrices were added to the feature set. The computational form of the matric properties expresses spatial changes in image intensity and the dynamic range of intensity; the matric being one-dimensional representation of summing the difference between the gray level of the pixel and the average gray level of the surrounding neighbors. These features tend to be more macroscopic than GLRL and GLSZ textures as they were developed to mimic human perception, including measures for coarseness, contrast, busyness, complexity, and strength of texture.

- Coarseness
- Contrast
- Busyness
- Complexity
- Strength

D.5. Size and Shape features

Volume, calculated as the number of voxels in the segmentation mask times the voxel dimensions; as it is rare for medical images to be isotropic in all three dimensions, it is possible this is a lower resolution than the other size metrics calculated only in the principle plane[80].

Two features novel to tumor classification applications were also included which take the area and diameter of the nodule and adjust them to the HU of the nodule; specifically calculating what the area and diameter of a nodule would be if it was entirely composed of water. Originally, these have been used in dose calculations of CT scans[94]. We included them in this expanded feature set as a potential supplement to traditional size measures in the case where segmentation variation can affect; with this measure, so long as the bulk segmentation is correct, the inclusion of a variable number of parenchymal border voxels has less effect on the overall measure.

- Volume
- H_2O Equivalent Diameter
- H_2O Equivalent Area

APPENDIX E: FEATURE SET REDUCTION

While the extraction of many features increases the confidence in finding features helpful to classification, it most likely includes some features that are highly related to each other. Highly correlated features in selection and classification can lead to model instability and decreased interpretability. To decrease the number of features with high correlations we proposed a feature set reduction approach where groups of highly inter-correlated features were condensed into a single representative feature; this approach was tested on two methods of reduction: k-medoids clustering[95] and principal component analysis (PCA)[157]. These tests were run using the original feature set on the original 50 cases (24 malignant, 26 benign); **Figure E.1** demonstrates the pairwise correlation between the extracted features on a heatmap[10]. Visualization of these tests is found in **Figure E.2**.To test the performance of the reduction method to preserve good-classifying features, the reduced feature set was pushed through a leave-one-out feed-forward (LOOFF) feature set selection method using an ANN with two hidden layer nodes, alpha = 1.716, beta = 0.6667, eeta = 0.03. Performance measures of AUC-ROC, sensitivity, and specificity are displayed in **Table E.1.**

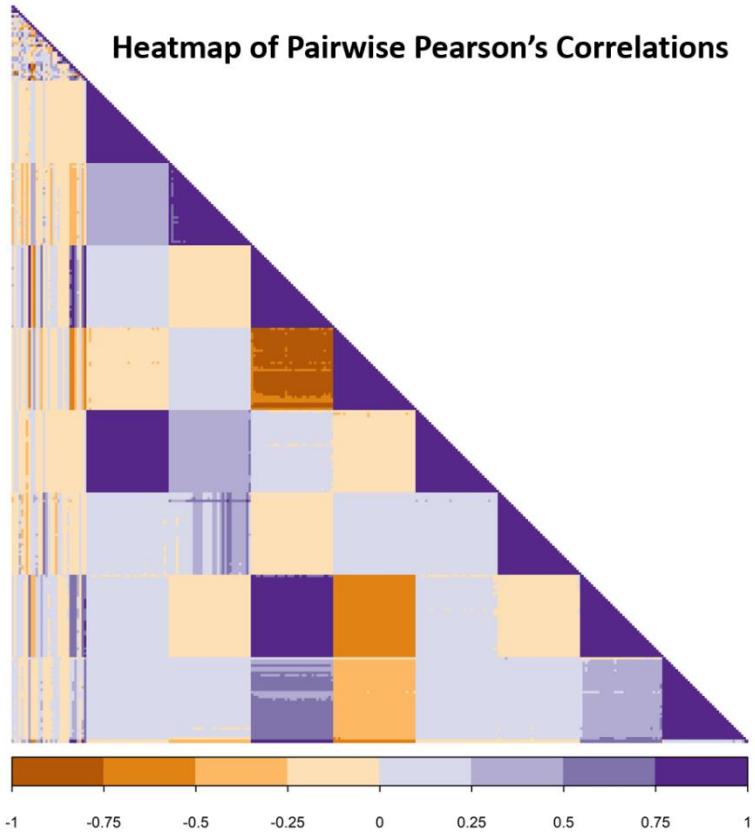

Figure E.1: Heatmap visualization of features' pairwise Pearson's correlations. Note large blocks of highly correlated features about the diagonal are groups of Law's texture energy measures.

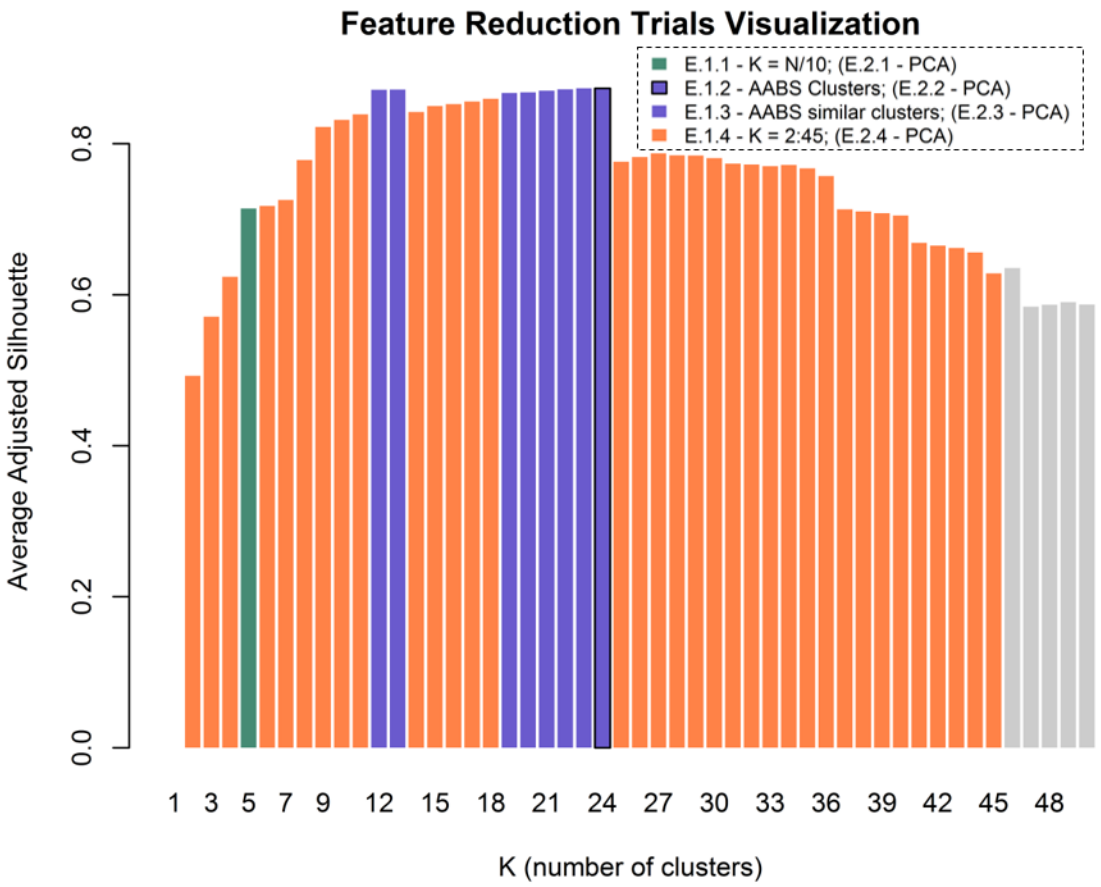

Figure E.2: Visualization of reduction tests on plot of k verses the average adjusted silhouette widths for k = 2:50. Definition of abbreviations: PCA – principle component analysis; AABS – average adjusted best silhouette

E.1. K-medoids Clustering

Clustering is a method of dimensionality reduction that seeks to group like-elements. The k-medoids clustering algorithm is a distance-based grouping procedure which attempts to produce optimal groups by maximizing a cluster's silhouette value by minimizing the sum of dissimilarities between objects and their medoids [95,96]. Each cluster is denoted by one silhouette signifying the proportion of objects within a cluster and in an intermediate clustering position; the greater the silhouette value, the more appropriate the clustering.

$$s(p) = \frac{nonmember_{least\ mean\ diss.}(p) - member_{mean\ diss.}(p)}{\max\langle nonmember_{least\ mean\ diss.}(p), member_{mean\ diss.}(p)\rangle} \rightarrow [-1,1]$$

[Equation E.1.1]

K-medoids is similar algorithmically to the popular k-means clustering except the objects are grouped around a representative object, termed the medoid, instead of the mean value. Here, we perform k-medoids clustering using a distance matrix composed of the pairwise correlations between features (**Figure E.1**). As the k-medoids algorithm requires the number for k to be known at the beginning, four

tests were devised to determine the k for feature reduction. These included setting k: (1) equal to the number of features to be selected, (2) equal to the k with the best average silhouette width, (3) equal to all k with similar silhouette widths to the best average silhouette width, and (4) equal to all k between 2 and N-5.

E.1.1. K = N/10 (Peduzzi limitation)

A baseline test of feature set reduction was run where the number of clusters equaled the heuristic limitation on maximum number of features set forth by Peduzzi, it was indicated that in a regression analysis model overfitting was generally prevented using a limitation of one predictor for every ten independent subjects[158].

Here, K was set to 50/10 = five clusters. The medoid features selected were all LTEM features: nodule mean-21, nodule variance-19, nodule kurtosis-31, parenchyma variance-25, and parenchyma skewness-20. The selection of LTEM features is not surprising as they make up a large proportion (272/386) of the full feature set and they tend to be highly correlated with each other which would drive silhouette optimization. The medoids of the resulting clusters were used in building the ANN; no additional feature selection was required or performed during this test.

E.1.2. K = best average silhouette width & LOOFF

A single cluster's silhouette is a goodness-of-fit measure for all points within a cluster. To find the best clustering among features, we calculated the average cluster silhouette produced by k-medoids with k from two to ninety-nine. For this feature dataset, the maximum average cluster silhouette was achieved with k = 24. The medoids of the twenty-four clusters were sent through the LOOFF selection resulting in an ANN built from the following five features: nodule minimum HU (IH), parenchyma entropy HU (IH), parenchyma variance-10 (LTEM), physical sphere variance (BASC), and standard deviation of slopes (BCRR).

E.1.3. Medoids with similar silhouette widths & LOOFF

From **Figure E.2**, it is clear that while k = 24 had the highest average silhouette measure there were other values of k which produced similarly high performing clusterings. It is possible then, these similar clusterings are just as good with very minor differences due to the dataset bias. For this test, we examined the results of using the medoids from k-clusterings with an average silhouette within 0.02 of the highest (0.8734).

E.1.4. K = 2:45 & LOOFF

From visual inspection of the full correlation matrix heatmap (**Figure E.1**), we estimated there could be a maximum of 45 clusters. For this test, we took any feature selected as a medoid for each of the

44 runs of k-medoids and ran it through LOOFF selection. Note, this provided the LOOFF selection pipeline with a similar number of features (63) to the Prior Approach (72).

E.2. Principle Component Analysis

Principle component analysis is a dimensionality reduction technique which finds the direction and spread of maximum variance and uses orthogonal transformations to convert the data points to a set of linearly uncorrelated variables. Visually, it can be thought of as fitting an N-D ellipsoid to the data set where each axis represents a principle component explaining a percentage of the variance in the dataset. The eigenvector, the direction of the principle component, can be used to transform data points into component scores. The component score is a combination of all the features included in the PCA and can be used as the representative feature. In this work, we have used principle component analysis on clusters determined by k-medoids clustering, meaning each principle component feature is the combination of all features in that cluster.

E.2.1-4. Application of PCA to clusters in E.1.1-E.1.4

Each of the four reduction tests performed in the k-medoids section were further reduced by using the principle component adjusted features for the clusters. These adjusted features were used in building the ANN using LOOFF set selection method.

- E.2.1 – Principle components of K = N/10 (Peduzzi limitation)
- E.2.2 – Principle components of K = AABS
- E.2.3 – Principle components of K with similar silhouettes to AABS
- E.2.4 – Principle components of K = 2:45

E.3. Selecting a Method of Feature Reduction

We selected AABS from k-medoids as the method for feature reduction as it obtained a high AUC-ROC and the highest sensitivity. In machine learning and medical tests in general, there is often a need for an unequal compromise between the sensitivity and the specificity of a tool depending on the specifics of the solution. In this case, we prioritize the case of over-diagnosing a benign nodule to the case of missing a malignancy. The risk of leaving a cancerous tumor could significantly allow for progression and possible metastasis before additional testing and treatment is sought while the risk of treating a benign nodule is secondary complications due to treatment or added testing. Both cases incur human costs in terms of unnecessary patient stress and decreased quality of care. From our prospective, we seek to error on the side of caution by prioritizing sensitivity measures over specificity; that being said, outlying cases will always exist, and classification performance should not be completely sacrificed to bring them into the correct classification.

Table E.1: ANN performance results from testing the feature set reduction methods on 50 subjects. Subject-level predictions and AUC-ROC were calculated from leave-one-out cross validation as described in section A.3.3. Sensitivity and specificity were calculated from the threshold described in section A.1.3.1.

Test	Reduction Method	Specifics	AUC-ROC	Sensitivity	Specificity
Prior	Statistical Significance	Section A.2.1	0.938	0.909	0.929
E.1.1	K-medoids	K = N/10	0.862	0.711	0.991
E.1.2	**K-medoids**	**AABS Clusters**	**0.939**	**0.944**	**0.934**
E.1.3	K-medoids	AABS similar clusters	0.929	0.925	0.931
E.1.4	K-medoids	K = 2:45	0.938	0.911	0.966
E.2.1	K-medoids + PCA	PCA of K = N/10	0.823	0.691	0.928
E.2.2	K-medoids + PCA	PCA of AABS	0.920	0.925	0.905
E.2.3	K-medoids + PCA	PCA of AABS similar clusters	0.920	0.916	0.910
E.2.4	K-medoids + PCA	PCA of K = 2:45	0.911	0.902	0.923

Definition of abbreviations: AUC-ROC – area-under-receiver-operator characteristics curve; N – number of subjects; AABS – average adjusted best silhouette; PCA – principle component analysis

As number of cases grows and number of features remains constant, we expect there to be a plateau in the optimal k which would represent the true nature of the features' interactions. The decision to use medoids as the feature reduction method has a two-fold advantage. Firstly, this method of clustering results in a representative feature, the medoid, which can provide insight into how features are affecting the classifier. K-medoids will not achieve a global minimum in all cases; however, the algorithm will always converge to a local minimum. To test the stability of the medoids in the dataset at a value of k, the method was run ten times with different seeded initializations to insure stable medoids. During the course of these investigations, we saw no variability in the medoid selection given the same dataset and value of k; therefore, we conclude the feature medoids are stable for most k.

APPENDIX F: FEATURE SET SELECTION

F.1. Methodology Development and Testing

The feature reduction methodology proposed in this dissertation (**Appendix E**) is performed independent of knowledge of classification. To determine the subset of features to use in classifier development, a feature set selection method is required. Methods have been proposed to assign rank to features either through forward selection of the best features or backwards elimination of least helpful features. The Prior Approach utilized LOOFF to determine the feature set for classifier development. While this was proven to be effective on a small cohort of 50 subjects and small number of features, it is computationally prohibitive in larger cohorts with more features available. We have methodically developed a method of feature set selection which is greatly improved in computational load while maintaining effectiveness on a large cohort of subjects. We tested six feature set selection methods on a cohort of 363 subjects (74 malignant, 289 benign) using an ANN with two hidden layer nodes, alpha = 1.716, beta = 0.6667, eeta = 0.03. Performance measures of AUC-ROC, sensitivity, and specificity are displayed in **Table F.1.**

F.1.1. Selecting K = N/10 (Peduzzi limitation)

Previously, in **E.1.1,** we experimented with the use of an alternate, but simple, form of feature selection in the wherein we selected the number of clusters for k-medoids to be equal to the desired number of selected features per the Peduzzi limitation. This process was repeated for this larger dataset with k = 36, this process required no additional computational load after feature reduction.

F.1.2. Selecting K = best average adjusted silhouette width (AABS)

In **E.1.2**, it was demonstrated using the k with the best average silhouette width (AABS) performed well when combined with the LOOFF. Here, we used the ranked individual cluster silhouette widths to select features to represent the set as a whole. Coincidently, k=36 was determined to be the optimal clustering based on average cluster silhouette with adjustment for solo clusters, therefore both methods of only using k-medoids clustering for feature selection yielded the same results as **F.1.1**.

F.1.3. Selecting N/10 with best Majority Votes from 10_x10_{fold} Cross Validation of K = best average silhouette width

kCV, as described in **A.3.2**, can be used to better assess variability in a method by portioning the dataset into randomized batches and altering which folds are used for training and validation. As dataset translatability is a very important, we investigated the change in medoid selection using 10 rounds of 10-fold cross validation (10_x10_{fold}kCV). This process yielded 60 unique medoid selected with fold-optimal medoid clustering k ranging from 32 to 41 (mean: 36.4±0.7). From the 60 unique medoids, we

implemented a majority votes selection process where features were ranked based on the number of times they appeared as a cluster medoid. The top 36 features were selected to build the ANN, the number of votes won (number of folds they were a medoid) by these features ranged from 389 to 71 (mean:120 ± 59).

F.1.4. Selecting N/10 using Information Theory and Random Forest Feature Importance Measures

A feedforward feature selection process was developed using two types of feature interrogator measures: information theory and random forest qualities. This type of feature set selection method requires an objective to maximize over the set of potential features. To determine the best objective function for the selection pipeline we investigated Monte Carlo method of different weightings of the eight measures (three information theory, five random forest) to produce **Equations F.1.4-6**.

Three information theory measures were extracted: mutual information of feature with class, interaction information within feature set, and mutual information of feature set with class [159]. The two specific mutual information terms indicate how much detail the feature or feature set can describe the class variables. The interaction information within a feature set describes the amount of shared information in a feature set, or in other terms, the sum of the intra-redundancy. In terms of feature selection, we want to limit the amount of information shared by the feature set and the next feature we are adding as shared information leads to less overall knowledge gained by the classifier by the addition of that feature. Random forest is an ensemble classifier built using many small decision trees which have been generated using random subsets of the features[160].

Five random forest importance measures were available: average Gini index, mean decrease in accuracy, importance on class, importance on malignant, and importance on benign. At its base, a decision tree attempts to create pure branches, or nodes; first by spitting the cases based on the feature which best reduces the entropy of the resulting branches. In a random forest, many short (highly pruned) trees are grown by repeatedly randomly selecting a subset of features to split on. While these trees will most likely not be pure at the final branches, the advantage of this method is it allows for many different features to be selected for nodes which can both increase the translatability of the model by decreasing model variance and provide measures of relative feature importance in classification.

$$Feature_{Selection} = arg_{max}(0.40 * MI_{class\&set} + 0.2 * II_{set} + 0.2 * RF_{class} + 0.15 * RF_{acc} + 0.05 * RF_{gini})$$

[Equation F.1.4]

Here, $\mathbf{MI_{class\&set}}$ is the mutual information between the set feature and classification, $\mathbf{II_{set}}$ is the interaction information between features in a set, $\mathbf{RF_{class}}$ is the random forest importance on class, $\mathbf{RF_{acc}}$ is the random forest mean decrease in accuracy and $\mathbf{RF_{gini}}$ is the random forest mean decrease in Gini index.

F.1.5. Selecting N/10 of Information Optimization Ranking

From the weighting schema, the best performing stable method using solely the three information theory measures was used to develop the Information Optimization (IO) method of feed-forward feature set selection. The IO feature selection method was applied on the 60 medoids from the reduction method of 10_x10_{fold}kCV. The objective function follows:

$$Feature_{Selection} = arg_{max}(0.675 * MI_{class\&set} + 0.225 * II_{set} + 0.1 * MI_{class\&feature})$$

[Equation F.1.5]

Here, $\mathbf{MI_{class\&feature}}$ is the mutual information between the feature and classification, $\mathbf{II_{set}}$ is the interaction information between features in a set, and $\mathbf{MI_{class\&set}}$ is the multi-mutual information between the feature set and classification.

F.1.6. Selecting N/10 of Random Forest Importance Optimization

From the weighting schema, the best performing stable method using solely the five random forest measures was used to develop the Random Forest Importance Optimization (RFIO) method of feed-forward feature set selection. The RFIO feature selection method was applied on the 60 medoids from the reduction method of 10_x10_{fold}kCV. The objective function follows:

$$Feature_{Selection} = arg_{max}(0.8 * RF_{class} + 0.15 * RF_{acc} + 0.05 * RF_{gini})$$

[Equation F.1.6]

Here, $\mathbf{RF_{class}}$ is the random forest importance on class, $\mathbf{RF_{acc}}$ is the random forest mean decrease in accuracy and $\mathbf{RF_{gini}}$ is the random forest mean decrease in Gini index.

F.2. Selecting a Method of Feature Selection

We selected IO as the method for feature set selection as it obtained a high AUC-ROC and the highest sensitivity (**Table F.1)**. The combination of k-medoids and IO can provide valuable insight into feature interaction and importance. For this dataset, it was coincidental that the best average of cluster's silhouette occurred at k equal to the maximum number of features to select; we expect that as a dataset grows this k would eventually plateau at an optimal value unique to the full feature set and the data source. We performed to additional checks, **F.2.1** and **F.2.2** to ensure we were selecting the superior method of combination feature set reduction (AABS) and feature set selection (IO).

Table F1: ANN performance results from testing the feature set reduction and methods on 363 subjects. Subject-level predictions and AUC-ROC were calculated from 10-fold cross validation as described in section A.3.2. Sensitivity and specificity were calculated from the threshold described in section A.1.3.1.

Test	Reduction Method	Selection Method	AUC-ROC	Sensitivity	Specificity
Prior	Statistical Significance	LOOFF	0.938	0.909	0.929
F.1.1	~	K=N/10	0.920	0.926	0.914
F.1.2	AABS	Silhouette Ranking	0.920	0.926	0.914
F.1.3	10_x10_{fold}kCV-AABS	Majority Votes	0.947	0.938	0.912
F.1.4	10_x10_{fold}kCV-AABS	IO+RFIO	0.950	0.962	0.901
F.1.5	**10_x10_{fold}kCV-AABS**	**IO**	**0.963**	**0.988**	**0.976**
F.1.6	10_x10_{fold}kCV-AABS	RFIO	0.942	0.952	0.895
F.2.1	~	IO	0.902	0.911	0.877
F.2.2	10_x10_{fold}kCV-AABS	IO Cluster-mate	0.958 to 0.963	0.961 to 0.988	0.954 to 0.978

Definition of abbreviations: AUC-ROC – area-under-receiver-operator characteristics curve; LOOFF – leave-one-subject-out feed-forward; N – number of subjects; AABS – average adjusted best silhouette; 10_x10_{fold}kCV – 10 rounds of 10-fold cross validation; IO – information optimization

F.2.1. Additional Consideration: Information Optimization Without Feature Set Reduction

Here, we consider the possibility that the feature set selection method may perform better without the feature set reduction set. To test this consideration, the IO method was run on the full un-reduced feature set and selected 36 features to build an ANN. This method showed a decrease in performance (AUC-ROC = 0.902), implying the benefit of feature set reduction of highly correlated features prior to feature set selection. This investigation has reinforced both the need for a sufficient reduction method and the need for feature selection beyond a first-cut reduction approach for classifier development.

F.2.2. Additional Consideration: Medoid Verses Cluster-mate Performance

As k-medoids selects a single representative feature for a cluster, there is the potential that performance of a classifier could be altered by the selection of specific representative features. The seminal case where this may be most altering would be in clusters with only two features. Currently, the method used for determining which feature is the medoid in clusters-of-two is the feature that is most dissimilar to the nearest cluster. On the final model of pipeline development, we did a test of substituting (one-at-a-time) in the cluster-mates for the clusters which contained only two features. This did not result in any significant difference in classification performance ($p > 0.05$) using Delong comparison of AUC-ROC curves. Raw prediction scores deviated by <0.001 to 0.042.

F.3. Set Size Maximum

In traditional machine learning algorithms using curated features, there is an art to deciding the number of features to use to develop the tool. This number needs to be large enough to capture the complexity of the problem's solution and the variability of the features between subjects with the same end classification. However, if the number is too large it can lead to more opportunities for overtraining

of the classifier, increased development time, and model complexity. The Peduzzi limitation for the number of features related to cases is one for every ten additional cases; this is a conservative method for restricting overfitting in the resultant model that has been adopted by many[158]. However, recently Vittinghoff et al. determined that decreasing ratio between the predictor and the number of unique cases to 1:5 did not lead to significant overtraining[98]. For the results presented in **Chapter 5: QIC-RATE** of this dissertation, the N/10 rule was relaxed to include up to N/5 features.

APPENDIX G: CLASSIFICATION

The final step in building the classification pipeline is choosing the type of classifier. There is a plethora of classification methods that have been developed[161]. Here, we selected classifiers from three distinct machine learning techniques: neural network, non-probabilistic linear discriminant, and decision tree. By selecting these different approaches, we can better assess a broader scope than if we selected three different subtypes of a method (i.e.: decision trees, random forest, conditional inference forests). Once an overall classification type is determined, tuning of the exact algorithm and parameters can take place.

G.1. Classification Methodology

G.1.1. Artificial Neural Network

An ANN is a classification method that consists of a collection of connected nodes linked together by weighted edges which loosely model the neurons in a brain. The training algorithm can be divided into two phases: propagation and weight update. The Prior Approach utilized an in-house ANN development script written in MatLab (Mathworks, Natick, MA)[11]. To summarize, this method was built using tanh sigmoid activation function. The default hyperparameters were as follows, single hidden layer, fully-connected, the hidden layer size = 2 nodes, alpha (activation) = 1.716, beta = 0.6667, eeta = 0.03. The script originally used non-seeded random initialization of weights; as such, each time the script is run the original weights are randomly initialized in a unique manner.

G.1.2. Support Vector Machine

From the category of non-probabilistic linear discriminants, an ensemble of 30 support vector machines (SVM-E) was built using gating of 8 experts. This method was run in R using the 'classyfire' package; the function 'cfBuild' was implemented with defaults .

G.1.3. Conditional Inference Forest

A forest ensemble of 30 conditional inference random trees (CIRF) testing 5 features at each node was implemented. This method was run in R using 'party' package; the function 'cforest' was implemented with defaults.

G.2. Selecting the Classifier

Based on 10-kCV results, tuning of exact algorithm and hyper-parameters was systematically performed. Of the three classification methods, SVM-E took the longest to run through 10-kCV taking approximately 5.25 hours, ENN run time was approximately 20.5 minutes, and CIRF had the shortest run time at approximately 4.75 minutes, when applied to the cohort. Results of the 10-kCV are shown in

Table G.1 below. The ANN classifier performed best in terms of accuracy, sensitivity, and AUC. It also had the highest specificity of the three, although SVM-E performed within the same standard deviation of 100 trials. The Youden threshold for the ANN classifier was a prediction output of 0.241. All cases with an ANN prediction value below were assigned benign classification and all cases equal to or greater than the cutoff was assigned malignant classification.

Table G.1: Performance results from testing the three classification methods on 363 subjects. Subject-level predictions and AUC-ROC were calculated from 10-fold cross validation as described in section A.3.2. Sensitivity and specificity were calculated from the threshold described in section A.1.3.1.

Section	Classification Method	AUC-ROC	Sensitivity	Specificity
G.1.1	**ANN**	**0.963**	**0.988**	**0.976**
G.1.2	SVM-E	0.864	0.602	0.959
G.1.3	CIRF	0.761	0.639	0.830

Definition of abbreviations: AUC-ROC – area-under-receiver-operator characteristics curve; ANN – artificial neural network; SVM-E – ensemble of support vector machines; CIRF – conditional inference random forest

In conclusion, the ANN achieved the best performance with an acceptable computational run time and therefore was selected to complete the pipeline. The most computationally dependent is the extraction of imaging features from the CT scan taking an average of 12 minutes per case. Once features are extracted from the ROI, the QIC-RATE method to build the final tool took an average of 7.2 minutes over ten runs. To run a new case through the complete pipeline (image segmentation, selected feature extraction, and classification prediction) took an average of 5 minutes for a validation cohort of ten cases.

G.3. Improvements to the ANN architecture

The architecture of ANNs in the Prior Approach was limited due to the classifier's use in LOOFF set selection, which necessitated a fast-building ANN framework as development was performed thousands of times during selection process. As such, the Prior Approach selected a single hidden layer architecture with hidden layer node size of two and hyper-parameters that were fixed. With the newly developed classifier development pipeline, feature set selection is independent of classifier development improving speed and negating the need for the repetitive classifier development of LOOFF set selection. Therefore, several ANN architecture tuning methods were employed to enhance the learning process and promote model stability.

G.3.1. Seeding of weight initialization

In the ANN architecture, there is the need to have initialized network weights before starting to train. Most often, these weights are randomly initialized through random number generators. A drawback to using random initialization is there is the potential for loss of repeatability of methods. To circumvent this and to ensure repeatability and therefore model development stability over training we have opted to used seeded random number generators. In seeding, a seed-point is selected and after the seed-point has

been called all sequential randomly generated numbers will be consistent. For example, calling a seed = 42 using the R seed function - set.sed(42) - followed by the random float number generator function - rnom(4) - produces the following sequence of numbers: [1.3709584, -0.5646982, 0.3631284, 0.6328626]

G.3.2. 10 rounds 10-kCV for each elemental ANN

Using 10 rounds of 10-kCV for each ANN checks on the influence of randomization in weight initialization (see also, **G.3.1**). Utilizing 10-kCV reduces the computational complexity over the traditional LOO while still providing a large proportion of the dataset for the training steps.

G.3.3. Increased complexity in elemental ANN architectures

We utilized a Monte Carlo simulation for the random selection hyperparameters, features in elemental ANNs. An array of 10,000 elements was generated by applying gaussian nose about the original fixed hyper-parameter values for each hyper-parameter. Architecture-specific parameters were also randomly selected in elemental ANNs including hidden layer size (2 to number of features), number of hidden layers (2 to 4), and the number of features applied.

G.3.4. Ensemble of Artificial Neural Networks

Stacked ensemble learners are a group of classifiers which come together to 'vote' or provide risk assessment as a panel of evaluators[162]. Ensemble learners built with sub-optimal or weak classifiers have been shown to typically outperform single classifiers on independent data as variability is more innately built into a group of classifiers. While the number of classifiers in an ensemble is up to the developer, often custom ensembles are built with 5 to 25 classifiers[163].