Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis

Joseph J. Foy
Samuel G. Armato, III
Hania A. Al-Hallaq

SPIE.

Joseph J. Foy, Samuel G. Armato III, Hania A. Al-Hallaq, "Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis," *J. Med. Imag.* **7**(1), 014504 (2020), doi: 10.1117/1.JMI.7.1.014504

Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis

Joseph J. Foy,[a] Samuel G. Armato III,[a,*] and Hania A. Al-Hallaq[b,*]
[a]The University of Chicago, Department of Radiology, Chicago, Illinois, United States
[b]The University of Chicago, Department of Radiation and Cellular Oncology, Chicago, Illinois, United States

Abstract

Purpose: While radiomics feature values can differ when extracted using different radiomics software, the effects of these variations when applied to a particular clinical task are currently unknown. The goal of our study was to use various radiomics software packages to classify patients with radiation pneumonitis (RP) and to quantify the variation in classification ability among packages.

Approach: A database of serial thoracic computed tomography scans was obtained from 105 patients with esophageal cancer. Patients were treated with radiation therapy (RT), resulting in 20 patients developing RP grade ≥2. Regions of interest (ROIs) were randomly placed in the lung volume of the pre-RT scan within high-dose regions (≥30 Gy), and corresponding ROIs were anatomically matched in the post-RT scan. Three radiomics packages were compared: A1 (in-house), IBEX v1.0 beta, and PyRadiomics v.2.0.0. Radiomics features robust to deformable registration and common among radiomics packages were calculated: four first-order and four gray-level co-occurrence matrix features. Differences in feature values between time points were calculated for each feature, and logistic regression was used in conjunction with analysis of variance to classify patients with and without RP ($p < 0.006$). Classification ability for each package was assessed using receiver operating characteristic (ROC) analysis and compared using the area under the ROC curve (AUC).

Results: Of the eight radiomics features, five were significantly correlated with RP status for all three packages, whereas one feature was not significantly correlated with RP for all three packages. The remaining two features differed in whether or not they were significantly associated with RP status among the packages. Seven of the eight features agreed among the packages in whether the AUC value was significantly >0.5.

Conclusions: Radiomics features extracted using different software packages can result in differences in classification ability.

© *2020 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.7.1.014504]

Keywords: radiomics; texture analysis; radiation pneumonitis; logistic regression; analysis of variance.

Paper 19215R received Aug. 16, 2019; accepted for publication Jan. 17, 2020; published online Feb. 21, 2020.

1 Introduction

Radiomics has shown great promise in classifying various diseases.[1–4] Recently, more advanced and accurate radiomics schemes have been developed, illustrating the promise that quantitative feature-based detection methods could play in a clinical setting. For example, studies have shown that radiomics techniques are capable of distinguishing between benign and malignant prostate lesions as well as offering additional information regarding the aggressiveness of

*Address all correspondence to Samuel G. Armato III, E-mail: s-armato@uchicago.edu; Hania A. Al-Hallaq, E-mail: hal-hallaq@radonc.uchicago.edu

2329-4302/2020/$28.00 © 2020 SPIE

the cancer.[5] Other studies showed promise in automatically segmenting lesion candidates in mammograms and classifying them as benign or malignant with high accuracy.[6]

Given the promising results that radiomics-based studies have reported, investigators have attempted to incorporate radiomics into their automated detection, diagnosis, and segmentation schemes.[4] The increased focus on this research has resulted in many institutions developing their own radiomics software, some of which has been made available so that others may use it. However, these packages have been used interchangeably without considering the inherent variations among them.[7] Variations across software packages can result from differences in algorithm implementation, image preprocessing and importation, or feature definitions. A number of studies have shown that differences in image acquisition parameters (e.g., voxel size, image reconstruction, or imaging system manufacturer) can cause large differences in radiomics feature values when extracted from any number of imaging modalities.[1,7–14]

Variations in feature values due to differences in image acquisition parameters may be exacerbated by computational differences embedded in the feature calculation software. Because it is often difficult to report a completely comprehensive outline of these methods, these studies are difficult to reproduce and validate, and the lack of reproducibility has slowed the clinical implementation of many promising radiomics-based detection and diagnosis schemes.

Some studies have illustrated the need for more standardized radiomics research.[7,11,15–19] Additional collaborations such as those put forth by the International Biomarker Standardization Initiative (IBSI) have attempted to standardize this workflow by providing recommendations regarding the various aspects of the feature-calculating process including image preprocessing, pixel interpolations, and feature definitions; however, this standardization has been applied to a relatively limited cohort of radiomics packages.[20,21]

Although it is known that radiomics feature values may differ when calculated at different institutions or with different radiomics software packages, the ultimate goal of such features is their use as imaging biomarkers.[7] Many freely available software packages have yet to become standardized to any particular reference, resulting in continued variability in radiomics research across institutions. Quantifying the differences in image-based radiomics feature values and understanding the sources of these variations are important; however, the effects of these variations when applied to a particular clinical task are currently unknown. Therefore, the purpose of this study was to use three radiomics packages to extract feature values from serial thoracic computed tomography (CT) scans of patients receiving radiation therapy (RT). Feature values were used to classify patients with and without radiation pneumonitis (RP), and classification ability associated with each radiomics package was compared.

2 Methods and Materials

2.1 *Medical Imaging Data*

A retrospective database of serial thoracic CT scans was acquired under institution review board approval from 105 patients receiving RT for esophageal cancer. Each patient underwent two high-resolution diagnostic CT scans, with the first scan acquired prior to treatment and the second acquired no more than 4 months after treatment (Table 1).[22,23]

A treatment planning scan and the associated dose map obtained from treatment planning were also acquired for each patient. Dose maps were generated using heterogeneity corrections using a Pinnacle (Philips Medical Systems, Andover, Massachusetts) treatment planning system for photon therapy or Eclipse (Variation Medical Systems, Palo Alto, California) treatment planning system for proton therapy. Patients were monitored for up to 6 months after treatment, and using all available documentation and imaging, RP status was determined through consensus of three clinicians using Common Toxicity Criteria for Adverse Effects, version 4. Each patient was assigned a binary value reflecting RP status: 1 for patients with RP (grade ≥ 2) or 0 for patients without RP (grade < 2), as shown in Fig. 1.

The patient- and treatment-specific variables shown in Table 1 and their association with RP were evaluated using the chi-squared test for nominal categorical variables and the

Table 1 Patient, treatment, and image characteristics represented as the number of patients belonging to that category and the relative number of patients belonging to that category represented as a percentage in parentheses.

	Parameter total	With RP	Without RP
No. of patients	105 (100%)	20 (19%)	85 (81%)
Gender			
Male	88 (84%)	17 (85%)	71 (84%)
Female	17 (16%)	3 (15%)	14 (16%)
Median age (range; years)	63 (27 to 81)	65 (48 to 81)	62 (27 to 79)
Smoking history			
Current	15 (14%)	2 (10%)	13 (15%)
Former	68 (65%)	13 (65%)	55 (65%)
Never	22 (21%)	5 (25%)	17 (20%)
Treatment modality			
IMRT	56 (53%)	9 (45%)	47 (55%)
3D-CRT	17 (16%)	4 (20%)	13 (15%)
Proton	32 (31%)	7 (35%)	25 (30%)
Treatment dose parameters			
Median prescribed dose (range; Gy)	50.4 (45 to 66)	50.4 (48.6 to 63)	50.4 (36 to 66)
Median no. of fractions (range)	28 (25 to 33)	28 (27 to 28)	28 (25 to 33)
Median MLD (range; Gy)	9.6 (2.5 to 18.7)	10.5 (2.9 to 15.2)	9.4 (2.5 to 18.7)
Median lung V20 (range; %)	16.6 (3.7 to 38.4	17.5 (4.8 to 31.3	16.3 (3.7 to 38.4
Median MRD (range; Gy)	37.9 (31.2 to 44.7)	38.2 (34.7 to 42.6)	37.8 (31.2 to 44.7)
Incidence of RP			
Grade 0	38 (36%)	—	—
Grade 1	47 (45%)	—	—
Grade 2	11 (10%)	—	—
Grade 3	5 (5%)	—	—
Grade 4	3 (3%)	—	—
Grade 5	1 (1%)	—	—

Note: MLD, mean lung dose; MRD, mean ROI dose.

Mann–Whitney U test for continuous variables. Significance was assessed at the 0.05 level after correcting for multiple comparisons using Bonferroni ($p < 0.006$).

The pre-RT, post-RT, and treatment planning CT scans were segmented using a semiautomated lung segmentation method and manually modified if necessary. Segmented post-RT scans were deformably registered to the pre-RT scans using the demons-based Plastimatch v1.5.12-beta software,[24] resulting in a vector map that matched corresponding anatomy between image acquisitions. The post-RT CT scans were not deformed themselves to preserve the texture of the captured tissue structure.

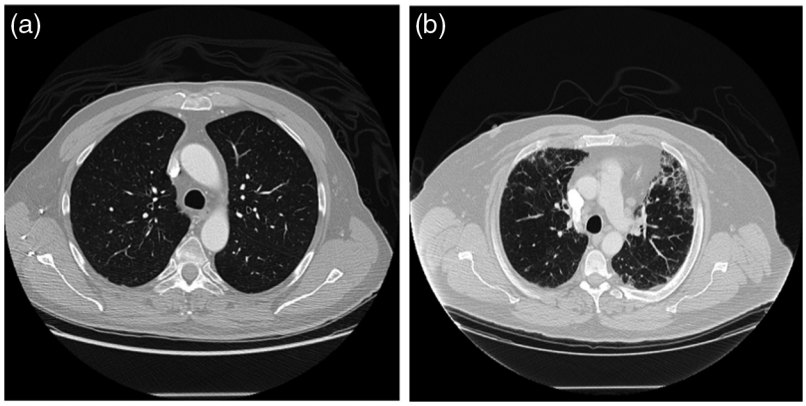


Fig. 1 CT scans illustrating the differences in texture for patients (a) without symptomatic RP (RP grade: 0) and (b) with symptomatic RP (RP grade: 5), which appears as higher-intensity pixels.

2.2 *Feature Calculation*

Regions of interest (ROIs) of 32×32 pixels in size (range in physical size: 20.0×20.0 to 31.3×31.3 mm^2) were randomly placed in the lung volume of the pre-RT scan for each patient without overlap and with a maximum of 10 ROIs placed in each axial slice. Corresponding ROIs were anatomically matched in the post-RT scan using the vector map obtained from deformable registration as has been outlined in our previous studies (Fig. 2).[22,23,25] In other words, the post-RT scan was not deformed thus preserving the texture of the image, but the vector map obtained from registration was used to anatomically match corresponding ROIs between time points.

In addition, the treatment planning scan was deformably registered to match the corresponding dose map to the pre-RT scan and enable calculation of the average planned radiation doses within each pre-RT ROI. Only ROIs placed in high-dose regions (≥30 Gy) were extracted, given

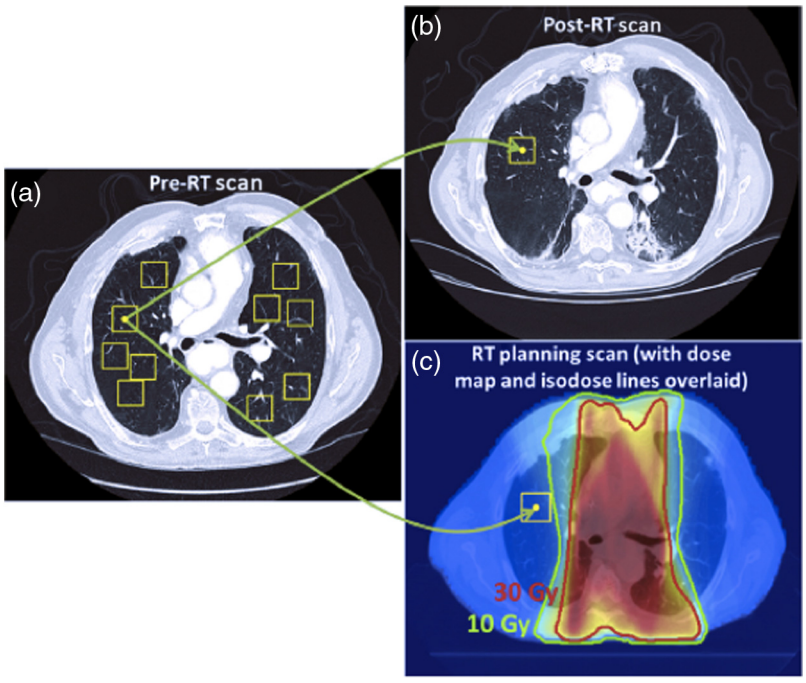


Fig. 2 (a) ROIs are randomly placed in the lung volume of the pre-RT scans, and (b) the vector map obtained from deformable registration anatomically matches ROIs in the post-RT scan. (c) The vector map obtained from deformably registering the treatment planning scan is used to match ROIs in the pre-RT scan to the anatomical locations in the treatment planning dose map, assigning a dose distribution to each ROI. Only ROIs placed in high-dose regions (≥30 Gy) were used. Reprinted with permission from Ref. 22.

Table 2 First- and second-order radiomics features common among all three packages and also robust to deformable registration.

First-order histogram features	Second-order GLCM features
Mean	Sum average
Median	Sum entropy
Min	Difference entropy
Entropy	Entropy

the previous results that showed that ROIs extracted from these regions were more predictive of RP development.[22,23]

Three radiomics packages were used for analysis: one in-house package (A1) (developed at The University of Chicago)[22] and two open-source packages, IBEX v1.0 beta (The University of Texas MD Anderson Cancer Center)[26] and PyRadiomics v.2.0.0.[27] These packages were used because they were the only packages that were freely available with source code at the initiation of this study. Each package could also be automated to process multiple ROIs in a single setting, and these packages had been cited in the literature. Each radiomics package was used to calculate all two-dimensional features common among the three packages that were also previously shown to be robust to deformable registration.[28,25] These features consisted of four first-order histogram features and four second-order gray-level co-occurrence matrix (GLCM) features (Table 2). While first-order features quantify the various attributes of the gray-level histogram of the pixel values, GLCM features characterize the spatial distribution of pixels within an ROI. The construction of the GLCMs prior to feature calculation can vary with the parameters used to describe these matrices, such as the gray-level limits, the number of gray levels, and the pixel binning.[29,30] Given that many radiomics studies do not report the GLCM parameters used, GLCM features were computed using the package-specific default parameters (Table 3). For every feature computed with each radiomics packages, a logistic regression model was constructed classifying patients with RP.

2.3 *Single-Feature Logistic Regression Modeling (M_{Avg})*

For each feature, the differences in feature values between image acquisitions were calculated using similar methods to those used in our previous studies[22,23]

$$\overline{\Delta \mathrm{FV}}_{F,S,p} = \frac{1}{N_p}\sum_{r=1}^{N_p}(\mathrm{FV}_{F,S,p,r}^{\mathrm{post-RT}} - \mathrm{FV}_{F,S,p,r}^{\mathrm{pre-RT}}), \tag{1}$$

where $\overline{\Delta \mathrm{FV}}_{F,S,p}$ is the average change in feature F calculated using software package S over all ROIs in patient p, N_p is the total number of ROIs placed in high-dose regions for patient p, and

Table 3 Package-specific default GLCM parameters. Pyradiomics designates 25 pixels for each bin of the GLCM, whereas A1 and IBEX designate the number of bins that compose the GLCM and the pixels are distributed evenly among the bins.

GLCM parameter	A1	IBEX	Pyradiomics
Gray-level limits	(−1500, 1500)	(Min, max)	(Min, max)
Number of gray levels	3001	(Max − min + 1)	Variable
Gray levels per bin	Variable	Variable	25
Number of directions	4	8	4

$\mathrm{FV}_{F,S,p,r}^{\mathrm{pre-RT}}$ and $\mathrm{FV}_{F,S,p,r}^{\mathrm{post-RT}}$ are the feature values computed in ROI r in the pre- and post-RT scans of patient p, respectively.

Using the dose map and the corresponding vector maps obtained from deformable registration, the mean ROI dose (MRD) for each patient was calculated using

$$\mathrm{MRD}_p = \frac{1}{N_p}\sum_{r=1}^{N_p} D_{p,r}, \tag{2}$$

where $D_{p,r}$ is the average planned radiation dose within the pre-RT ROI r of patient p, and N_p is the total number of ROIs placed in the high-dose region of patient p. This resulted in one value of MRD for each patient.

A logistic regression model was developed using the pROC package in R v3.3.3 classifying patients with and without RP. Previous studies have reported that treatment-specific dose parameters are correlated with RP development; however, these results vary across institutions.[31,32] Therefore, logistic regression models were constructed using MRD alone, and individual features were added to this model to determine whether the addition of these features significantly improved classification ability using receiver operating characteristic (ROC) analysis, with the area under the ROC curve (AUC) as the performance assessment metric. Individual radiomics features calculated using each package were then added to the logistic regression model

$$\mathrm{RP} \sim \mathrm{MRD} + \overline{\Delta\mathrm{FV}}_{F,S}, \tag{3}$$

where RP is a binary classifier indicating whether or not a patient develops RP (grade ≥ 2), MRD is the mean ROI dose for each patient, and $\overline{\Delta\mathrm{FV}}_{F,S}$ is the mean change in each of the eight radiomics features (F) calculated using each of the three radiomics software packages (S). Models of this form using a single feature with changes in feature values averaged over all ROIs for each patient are referred to as M_{Avg} for clarity. Analysis of variance (ANOVA) was used with chi-squared tests to determine whether the addition of each feature to the logistic regression significantly improved classification ability over using MRD by itself. The parameter MRD was also replaced in the regression model with the mean lung dose (MLD) and the relative volume of the lung that received at least 20 Gy (V20) to determine whether these parameters resulted in a different set of features that were significantly associated with RP. Significance was assessed at the $\alpha = 0.05$ level after correcting for multiple comparisons using Bonferroni test ($p < 0.002$).

During logistic regression, patient data were randomly sampled so that 50% of the patients were used for training the regression model, whereas the remaining 50% were reserved for validation. Sampling was performed to maintain the ratio of RP-positive to RP-negative patients in both the training and validation sets. Random sampling was performed 1000 times, and an AUC value was calculated for each iteration, resulting in a mean AUC value across iterations along with the corresponding 95% confidence intervals.

2.4 *Multifeature Logistic Regression Modeling*

To determine whether combinations of features significantly improved classification ability and whether feature combinations differed among packages, an additional feature was added to the logistic regression model using the following equation:

$$\mathrm{RP} \sim \mathrm{MRD} + \overline{\Delta\mathrm{FV}}_{F1,S} + \overline{\Delta\mathrm{FV}}_{F2,S}, \tag{4}$$

where the subscripts $F1$ and $F2$ refer to the first and second features added to the model, respectively, and S refers to the software package used to calculate these features, similar to that shown in Eq. (3). Models were first created using each of the eight individual features, and the seven remaining features were added to each of these models, resulting in a total of 56 total feature combinations. Significance was assessed at the 0.05 level after correcting for the 56 comparisons per package ($p < 0.0009$). Because of the limited number of RP-positive patients in this dataset, the potential for overfitting was a concern. Therefore, the Akaike information criterion (AIC)

was used to assess the relative quality of the models when the first and second features were included in the model compared with when only the first feature was included. The AIC quantifies model quality by balancing the potential for improved goodness of fit due to additional feature inclusion with the deficit introduced by the potential for overfitting.[33] The AIC was calculated using the following equation:

$$\mathrm{AIC}_{F1,F2,S} = -2(\log-\mathrm{likelihood}) + 2(n_{\mathrm{par}}), \tag{5}$$

where $\mathrm{AIC}_{F1,F2,S}$ is the AIC value when feature $F2$ was added to a model that included the first feature, $F1$ and MRD when both features were calculated using software package S. The log-likelihood reflects how well the model fits the data and n_{par} is the number of parameters. The absolute value of the AIC is arbitrary but feature combinations that result in smaller AIC values compared with models including one feature correspond to models of greater relative quality; in other words, the potential for overfitting due to the increased number of parameters in the regression is outweighed by the improved model fit reflected by the log-likelihood.

2.5 *Single-Feature Individual ROI Pair Modeling (M_{Ind})*

Averaging the difference in feature values over all ROIs for each patient could potentially dampen the impact of ROI pairs that demonstrate larger texture differences and therefore may be less indicative of RP development. Therefore, an additional set of models was constructed using each ROI pair as a distinct analyzable unit when training the logistic regression model (Fig. 3).

$$\Delta\mathrm{FV}_{F,S,p,r} = \mathrm{FV}^{\mathrm{post-RT}}_{F,S,p,r} - \mathrm{FV}^{\mathrm{pre-RT}}_{F,S,p,r}. \tag{6}$$

Similar to what is shown in Eq. (1), $\mathrm{FV}^{\mathrm{post-RT}}_{F,S,p,r}$ and $\mathrm{FV}^{\mathrm{pre-RT}}_{F,S,p,r}$ are the values for feature F computed using software S in ROI r in the pre- and post-RT scans of patient p, respectively. For these models, $\Delta\mathrm{FV}_{F,S,p,r}$ was included in the logistic regression for each feature along with the mean dose to the corresponding ROIs [$D_{p,r}$ in Eq. (2)], resulting in a total of 4474 analyzable units instead of the 105 units used previously.

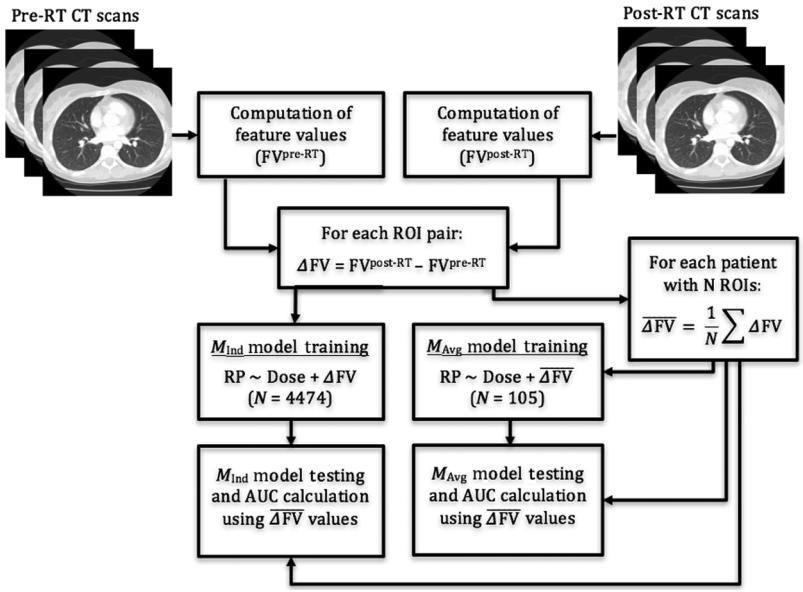


Fig. 3 Flowchart depicting regression models trained using individual ROI pairs (M_{Ind} models) as well as averages over ROI pairs for each patient (M_{Avg} models). Both models are validated using $\overline{\Delta\mathrm{FV}}_{F,S,p}$ values but were trained differently.

$$\mathrm{RP} \sim \mathrm{D} + \Delta\mathrm{FV}_{F,S}. \tag{7}$$

Models using a single feature (F) and individual ROI pairs are referred to as M_{Ind} models. ANOVA was used to determine whether the addition of each feature to logistic regression significantly improved model fit over using just the mean dose to the ROIs. During the validation process, both M_{Ind} and M_{Avg} models were assessed using the changes in feature value averaged over all ROIs for each patient [$\overline{\Delta\mathrm{FV}}_{F,S,p}$ from Eq. (1)] so that both models were assessed using the same patient data while also validating the M_{Ind} models using uncorrelated $\overline{\Delta\mathrm{FV}}_{F,S,p}$ values. In other words, ROIs from the same patient cannot be used to validate classification models, given that ROIs extracted from the same patient scans are likely to have feature values that are correlated with one another. In addition, if each patient contributes a different number of ROIs, this would skew the classification ability of the model. It should be noted that for both models, patients were split into training and validation sets, such that ROIs from the same patient were not used in both training and validation sets during the same sampling iteration. Half of the patients were sampled for training and all individual ROIs corresponding to those patients were used for training while the remaining ROIs were averaged for each patient and used for validation, resulting in one value of $\overline{\Delta\mathrm{FV}}_{F,S,p}$ for each feature for a given patient. Sampling, training, and validation were performed in the same way as described in Sec. 2.3. M_{Ind} and M_{Avg} models were compared using Vuong's closeness test of non-nested model comparison.[34]

3 Results

Patient- and treatment-specific variables shown in Table 1 were not significantly correlated with RP status.

3.1 *Single-Feature Logistic Regression Modeling*

Four first-order gray-level features and four second-order GLCM features were calculated using three different radiomics packages and using ROIs extracted from the CT scans. For each feature, $\overline{\Delta\mathrm{FV}}_{F,S,p}$ values were added to a logistic regression model including only MRD (M_{Avg} models) to determine whether the addition of the feature significantly improved the classification ability of that model (Table 4). Packages typically agreed regarding the features that were significantly correlated with RP other than GLCM difference entropy and GLCM entropy, which only showed significant correlation for features extracted using Pyradiomics. Dosimetric parameters alone

Table 4 The p values indicating whether the addition of a particular feature significantly improved classification ability over using the MRD alone. Features that were considered significantly associated with RP did not change when the V20 or MLD was used in the regression instead of the MRD. Each regression model was trained using averages in changes in feature values over all ROIs for each patient (M_{Avg}).

Feature	A1	IBEX	Pyradiomics
Mean	$p < 0.002$	$p < 0.002$	$p < 0.002$
Min	0.593	0.133	0.593
Median	$p < 0.002$	$p < 0.002$	$p < 0.002$
Entropy	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM sum average	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM sum entropy	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM difference entropy	0.008	0.010	$p < 0.002$
GLCM entropy	0.222	0.947	$p < 0.002$

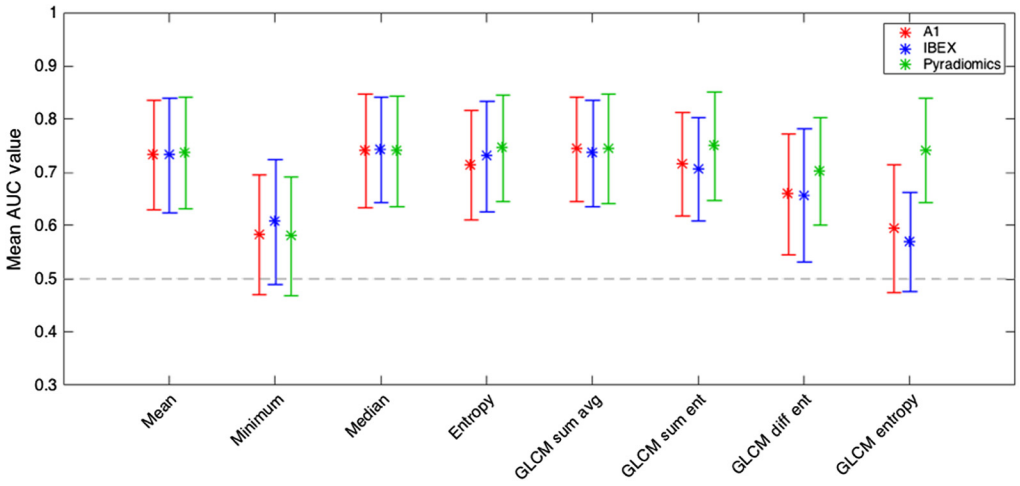


Fig. 4 Mean AUC values along with the corresponding 95% confidence intervals for eight features used to train M_{Avg} models. While packages A1 and IBEX were not significantly >0.5 for GLCM entropy, Pyradiomics was.

(MRD, V20, and MLD) were not found to be significantly associated with RP ($p = 0.19$, 0.24, and 0.10, respectively). Inclusion of any of the three dosimetric parameters into a regression model did not change which features were significantly correlated with RP as shown in Table 4.

The mean AUC values along with the 95% confidence intervals for each of the eight features are shown in Fig. 4.

3.2 *Multifeature Logistic Regression Modeling*

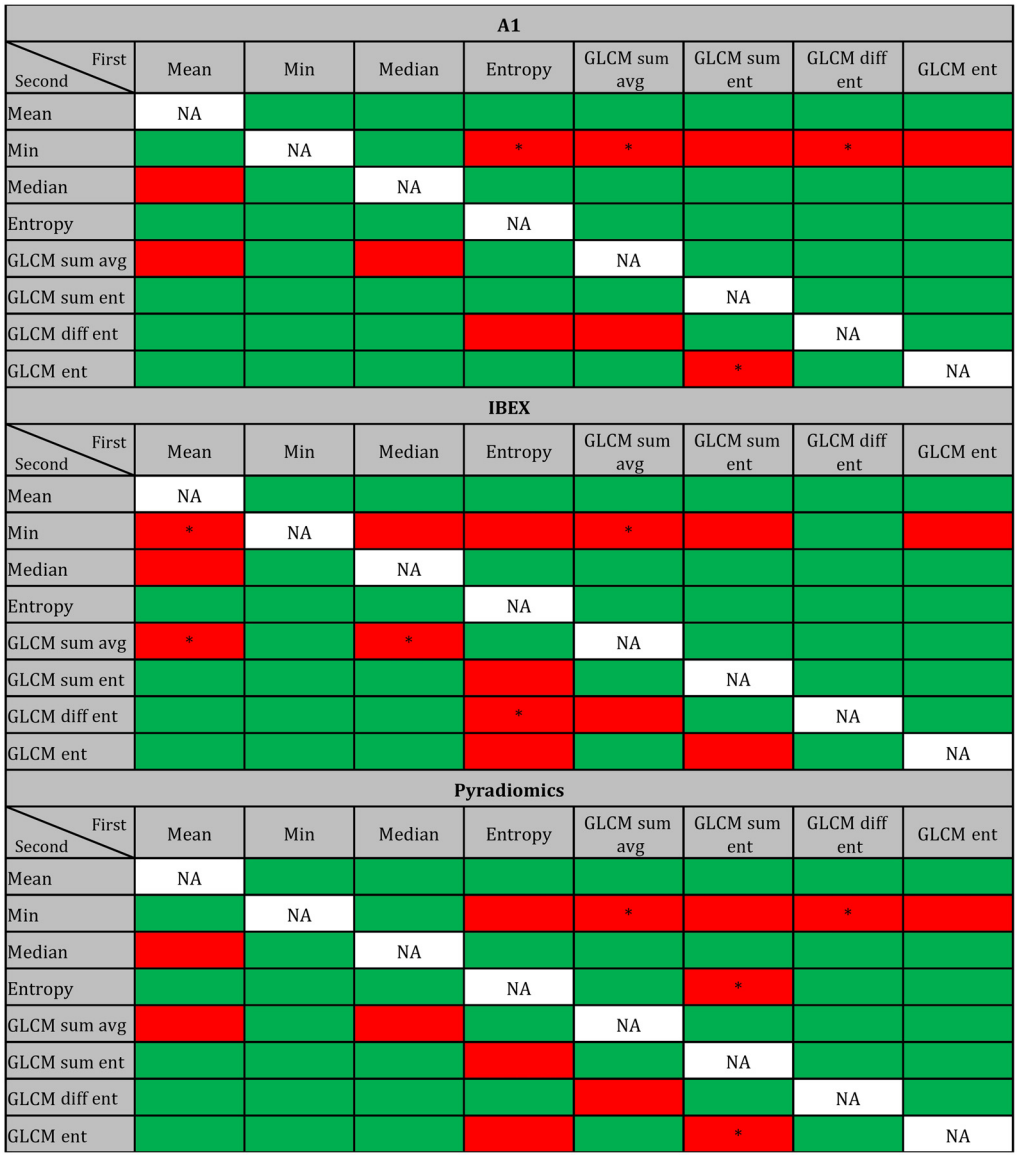
Figure 5 illustrates feature combinations in green that significantly improve model fit over using the first feature and the MRD alone ($p < 0.0009$). Cells in red illustrate features that did not improve model fit when added to the first feature in the model.

The addition of a second feature significantly improved model fit for most features for all three packages. Feature combinations that significantly improved model fit tended to agree among packages. Of the 56 feature combinations for each package, 40 (71%) combinations resulted in significant improvement (green) in model fit for all three packages, 9 (16%) feature combinations did not result in significant improvement (red) in all three packages, and the remaining 7 (13%) combinations differed among packages. For example, when GLCM sum entropy or GLCM difference entropy was added to a model already containing first-order entropy, the effect on the model fit differed among the three packages as shown in Fig. 5. There were multiple feature combinations that disagreed among packages when first-order entropy was the first feature included in the model. When adding GLCM difference entropy to a model using first-order entropy, the additional feature significantly improves model fit when features are calculated using Pyradiomics but not for packages A1 or IBEX. When assessing model quality based on AIC, all feature combinations shown in green had AIC values that were lower (corresponding to higher model quality) than when just the first feature and MRD were used in the model.

3.3 *Individual ROI Pair Logistic Regression Modeling*

When individual ROI pairs ($\Delta\text{FV}_{F,S,p,r}$) were used to train the logistic regression models instead of the average of the differences in feature values across ROIs for each patient ($\overline{\Delta\text{FV}}_{F,S,p}$), a greater number of features were significantly correlated with RP status as shown in Table 5. The bolded p values are features that were not considered to be correlated with RP status for the M_{Avg} models.

The mean AUC values and the corresponding 95% confidence intervals are shown in Fig. 6. The same trends in mean AUC values arise for M_{Ind} and M_{Avg} models, but the mean AUC values



A1

First / Second	Mean	Min	Median	Entropy	GLCM sum avg	GLCM sum ent	GLCM diff ent	GLCM ent
Mean	NA							
Min		NA		*	*		*	
Median			NA					
Entropy				NA				
GLCM sum avg					NA			
GLCM sum ent						NA		
GLCM diff ent							NA	
GLCM ent						*		NA

IBEX

First / Second	Mean	Min	Median	Entropy	GLCM sum avg	GLCM sum ent	GLCM diff ent	GLCM ent
Mean	NA							
Min	*	NA			*			
Median			NA					
Entropy				NA				
GLCM sum avg	*		*		NA			
GLCM sum ent						NA		
GLCM diff ent				*			NA	
GLCM ent								NA

Pyradiomics

First / Second	Mean	Min	Median	Entropy	GLCM sum avg	GLCM sum ent	GLCM diff ent	GLCM ent
Mean	NA							
Min		NA			*		*	
Median			NA					
Entropy				NA		*		
GLCM sum avg					NA			
GLCM sum ent						NA		
GLCM diff ent							NA	
GLCM ent						*		NA

Fig. 5 Green cells indicate the addition of a second feature in logistic regression that significantly improved model fit over using the first feature and MRD alone when features were calculated using package A1, IBEX, and Pyradiomics. Columns correspond to the first feature included in the model, and rows correspond to the second feature added to the model. Significance was assessed at the 0.05 level after correcting for the 56 different comparisons per package ($p < 0.0009$). Cells labeled with an asterisk reflect feature combinations resulting in greater AIC values (lower model quality) than when only the first feature was included in the model. Each regression model was trained using averages of changes in feature values over all ROIs for each patient (M_{Avg}).

were slightly greater when trained using M_{Ind} models rather than M_{Avg} models for all packages and for all features besides first-order and GLCM entropy from A1. However, the models constructed using the two methods (i.e., M_{avg} versus M_{ind}) were not significantly distinguishable based on Vuong's closeness test.

4 Discussion

This study demonstrated the variability in classification ability when radiomics features were computed using three radiomics software packages and applied to a clinically relevant classification task. Previous studies have localized this variability in software to discrepancies in

Table 5 The p values indicating whether the addition of a particular feature significantly improved classification ability over using the MRD alone when individual ROI pairs were used in the training of each model (M_{Ind}). During validation, averages of changes in feature values were used.

Feature	A1	IBEX	Pyradiomics
Mean	$p < 0.002$	$p < 0.002$	$p < 0.002$
Min	0.024	$\mathbf{p < 0.002}$	0.024
Median	$p < 0.002$	$p < 0.002$	$p < 0.002$
Entropy	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM sum average	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM sum entropy	$p < 0.002$	$p < 0.002$	$p < 0.002$
GLCM difference entropy	$\mathbf{p < 0.002}$	$\mathbf{p < 0.002}$	$p < 0.002$
GLCM entropy	$\mathbf{p < 0.002}$	$\mathbf{p < 0.002}$	$p < 0.002$

Note: Bolded p-values indicate features that were considered significantly correlated with RP for M_{Ind} models but not for M_{Avg} models.

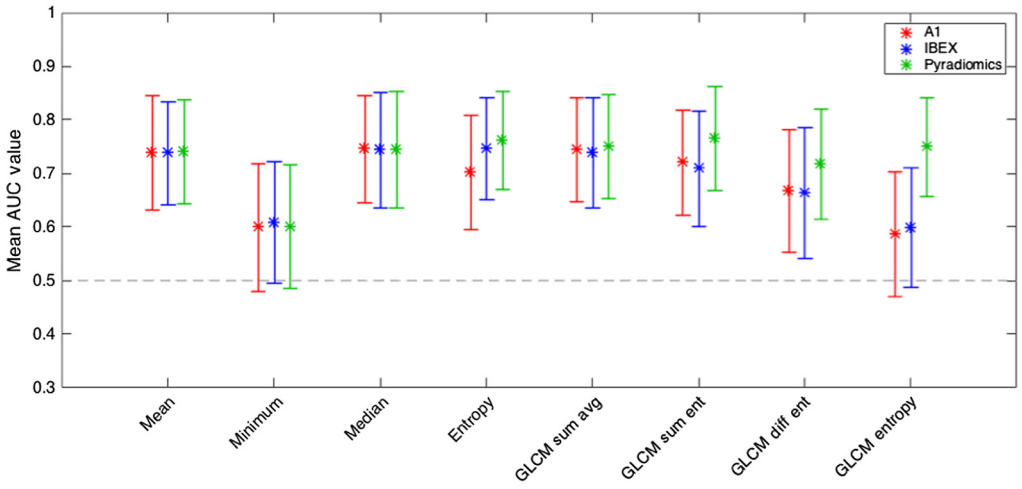


Fig. 6 Mean AUC values along with the corresponding 95% confidence intervals for eight features when individual ROI pairs were used in model training.

a number of aspects in the feature calculation pipeline including differences in image importation and preprocessing, algorithm implementation, and feature-specific calculation parameters.[7,35] Although many first-order features showed agreement in correlation with RP status, two of the four GLCM features disagreed among the packages (Table 4). A similar trend was shown in the mean AUC values plotted in Fig. 4; GLCM entropy was not significantly >0.5 for packages A1 and IBEX but was >0.5 for Pyradiomics. Our previous studies have reported that GLCM feature values are dependent on the parameters used to construct the matrices prior to feature calculation, and use of the default GLCM parameters unique to each package can alter the resultant feature values.[7] When the GLCM parameters were made to be consistent across the three packages [gray-level limits: (min, max); number of gray levels: 64; number of directions: 8], all four GLCM features were significantly correlated with RP status for all three packages. In addition, all four GLCM features had AUC values that were significantly >0.5 for all three packages, indicating that the parameters used to calculate various radiomics features can greatly affect the agreement in classification ability among software packages. While modifying these parameters may result in greater agreement among radiomics packages, many radiomics-based studies do not report complete definitions of each feature. Consequently, independent investigators cannot accurately recalculate these features, so package-specific default GLCM parameters are often

used. When applied to the clinical task of classifying patients with RP, the effects of these discrepancies are still noticeable: mean AUC values agreed much more among packages for first-order features compared with GLCM features. Consequently, when investigators attempt to reproduce, validate, or advance radiomics studies reported in the literature using different radiomics software packages, they may identify a different set of correlated features or achieve different levels of classification ability.

A common method of increasing AUC value for many radiomics-based classification schemes is to combine features during model construction, if the dataset is large enough. In this study, pairwise combinations of features were used in logistic regression, and the addition of a second feature significantly improved the model fit over using the first feature alone for the majority of combinations. While features that improved model fit over dose alone agreed among packages, feature combinations varied from package to package as shown in Fig. 5. Furthermore, feature combinations that significantly improved model fit differed more for GLCM features than for first-order features because of the increased complexity of GLCM feature calculation and the potential for larger discrepancies in GLCM feature values computed using different software. Moreover, feature combinations that decreased the model quality based on AIC (Fig. 5) showed very little agreement among packages, further illustrating that radiomics studies reporting promising features or combinations of features may not translate to other institutions using different radiomics packages.

When individual ROI pairs were used in model training (M_{Ind} models) instead of averages over ROI pairs (M_{Avg} models), more features were significantly correlated with RP status, and seven out of eight features agreed among packages. The discrepancies in correlated features between M_{Ind} and M_{Avg} models illustrate that differences in model implementation can affect the library of features capable of accurately classifying patients into various disease states. Despite the set of correlated features differing between the M_{Ind} and M_{Avg} models, the predictive ability of the two models was not significantly distinguishable based on Vuong's closeness test for any feature computed using any package; however, when classifying other diseases or using larger patient databases, M_{Ind} models may be more sensitive to changes in tissue structure and result in greater classification ability.

Previously conducted studies have recognized and attempted to address the need for greater standardization in radiomics research. The IBSI has combined the efforts of 19 institutions to compile a comprehensive manual of feature definitions and image-processing protocols. These institutions used their respective radiomics software to calculate various features on a small digital phantom and subsequently on a CT scan from a patient with lung cancer. Each software package was then iteratively modified and was considered "standardized" if at least half of them achieved the same feature values. Through this process, 99.4% of features agreed when computed on the digital phantom and 96.4% of features agreed when computed on the CT scan;[20,21] however, complete agreement among institutions was not achieved. This study illustrates that a broader standardization initiative must be conducted that includes both in-house and open-source radiomics packages while also considering the effects of standardization on various clinically relevant tasks (e.g., classification or segmentation task). While this study does not aim to offer recommendations outside of those reported by the IBSI, it supports the notion that greater harmonization of radiomics research must be achieved to obtain greater clinical implementation.

Our study also illustrates that the variation in feature values extracted from the same images does not necessarily result in variation in classification ability. Our prior study found significant differences in feature values among radiomics packages for all features extracted from head and neck tumors in CT scans and mammography scans.[7] In this study, all feature values also reflected significant differences across packages, but the variability in feature values was much greater than the variability in AUC values when assessed using the coefficient of variation (COV). For example, when the mean of GLCM sum average is calculated across all ROIs for each package, the COV among the three packages for the pre- and post-RT ROIs were 1.456 and 1.403, respectively. In comparison, the COV for the AUC values among the three packages was 0.006, indicating much less variability in AUC values than feature values across packages.

Despite the relatively large differences in the feature values themselves, when assessing which features were significantly correlated with RP (Table 4), six of the eight features agreed among packages. This is in part due to the inherent nature of delta radiomics schemes: variations

in radiomics feature values introduced by each software package because of differences in image preprocessing or algorithm implementation can be negated when assessing the changes in radiomics features over time. Therefore, it may not be sufficient to quantify only the differences in radiomics features but also how these differences translate into clinical practice when applied to a particular task.

This study contained a number of limitations in its methodology and areas that could be improved upon in future studies. First, only three radiomics packages were used in this investigation. Future studies could include additional radiomics packages or combinations of packages to allow for evaluation of a greater number of features. This study was limited to only the eight first-order and GLCM features that were common among the three packages; however, additional feature categories, such as fractal, Fourier, or gray-level run-length matrix features, may display variations in classification ability when computed using different software packages. Investigating these additional feature categories may illustrate additional areas of discrepancy among radiomics software, which will subsequently aid in the overarching aim of standardizing the radiomics workflow to make future research studies more reproducible and more translatable.[20]

Additional studies could also investigate the variability in classification ability when applied to different diseases and imaging modalities. Previous studies have shown that the degree of variability in raw feature values among packages can differ depending on the tissues being analyzed and the modalities used to image these tissues.[7] Radiomics packages are often developed and designed to analyze a particular range of pixel values or particular disease or tissue type. When these packages are used to analyze images beyond those for which they were designed for, the subsequent feature values may be meaningless and not reflective of the true texture. Therefore, it may be beneficial for future studies to analyze images outside of normal lung CT to determine if classification ability differs when applied to a different clinical task.

5 Conclusion

This study investigated the variability in classification ability among three radiomics packages for distinguishing patients with and without RP. When assessing which features were significantly correlated with RP, first-order features reflected greater agreement among packages, whereas GLCM features reflected greater variation. When additional features were added to the logistic regression models, feature combinations that improved classification ability over using the first feature alone also differed among packages. Initiatives have worked toward standardizing the radiomics workflow across institutions; however, the present findings indicate that additional effort must be put toward harmonizing radiomics research to achieve greater clinical implementation of their results.

Disclosures

This work was supported in part by the Cancer Center Support, National Cancer Institute (Grant No. P30CA014599) and the Jules J. Reingold Fellowship Endowment. S.G.A. and H.A.A. received royalties and licensing fees for computer-aided diagnosis technology through The University of Chicago.

Acknowledgments

The authors would like to thank Edward Castillo, PhD; Richard Castillo, PhD; and Thomas Guerrero, MD, PhD for facilitating data sharing with MD Anderson. They would also like to thank Kristen Wroblewski, Department of Health Studies, The University of Chicago, for her guidance in statistical analysis.

References

1. R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology* **278**(2), 563–577 (2016).

2. S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Phys. Med. Biol.* **61**(13), R150–R166 (2016).
3. H. J. Aerts et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.* **5**, 4006 (2014).
4. M. L. Giger, H. P. Chan, and J. Boone, "Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM," *Med. Phys.* **35**(12), 5799–5820 (2008).
5. A Wibmer et al., "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores," *Eur. Radiol.* **25**(10), 2840–2850 (2015).
6. S. G. Sapate et al., "Radiomics based detection and characterization of suspicious lesions on full field digital mammograms," *Comp. Methods Prog. Biomed.* **163**, 1–20 (2018).
7. J. J. Foy et al., "Variation in algorithm implementation across radiomics software," *J. Med. Imaging* **5**(4), 044505 (2018).
8. J. A. Oliver et al., "Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer," *Transl. Oncol.* **8**(6), 524–534 (2015).
9. P. E. Galavis et al., "Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters," *Acta Oncol.* **49**(7), 1012–1016 (2010).
10. N. M. Cheng, Y. H. Fang, and T. C. Yen, "The promise and limits of PET texture analysis," *Ann. Nucl. Med.* **27**(9), 867–869 (2013).
11. R. T. Leijenaar et al., "The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis," *Sci. Rep.* **5**(5), 11075 (2015).
12. M. Shafiq-Ul-Hassan et al., "Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels," *Med. Phys.* **44**(3), 1050–1062 (2017).
13. D. Mackin et al., "Measuring computed tomography scanner variability of radiomics features," *Invest Radiol.* **50**(11), 757–765 (2015).
14. K. R. Mendel et al., "Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers' systems," *J. Med. Imaging* **5**(1), 011002 (2017).
15. M. Sollini et al., "PET radiomics in NSCLC: state of the art and a proposal for harmonization of methodology," *Sci. Rep.* **7**(1), 358 (2017).
16. M. J. Nyflot et al., "Quantitative radiomics: impact of stochastic effects on textural feature analysis implies needs for standards," *J. Med. Imaging* **2**(4), 041002 (2015).
17. M. Hatt et al., "Characterization of PET/CT images using texture analysis: the past, the present... any future?" *Eur. J. Nucl. Med. Mol. Imaging* **44**(1), 151–165 (2017).
18. A. Traverso et al., "Repeatability and reproducibility of radiomics features: a systemic review," *Int. J. Radiat. Oncol. Biol. Phys.* **102**(4), 1143–1158 (2018).
19. J. Kalpathy-Cramer et al., "Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features," *Tomography* **2**(4), 430–437 (2016).
20. A. Zwanenburg et al., "Image biomarker standardisation initiative," arXiv:1612.07003 (2016).
21. A. Zwanenburg, "EP-1677: multicentre initiative for standardisation of image biomarkers [abstract]," *Radiother. Oncol.* **123**(Suppl.), S914–S915 (2017).
22. A. Cunliffe et al., "Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development," *Int. J. Radiat. Oncol. Biol. Phys.* **91**(5), 1048–1056 (2015).
23. G. J. Anthony et al., "Incorporation of pre-therapy ^{18}F-FDG uptake data with CT texture features into a radiomics model for radiation pneumonitis diagnosis," *Med. Phys.* **44**(7), 3686–3694 (2017).
24. G. C. Sharp et al., "GPU-based streaming architectures for fast cone-beam CT image reconstruction and demons deformable registration," *Phys. Med. Biol.* **52**(19), 5771–5783 (2007).
25. A. R. Cunliffe et al., "Lung texture in serial thoracic CT scans: registration-based methods to compare anatomically matched regions," *Med. Phys.* **40**(6), 061906 (2013).
26. L. Zhang et al., "IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics," *Med. Phys.* **42**(3), 1341–1353 (2015).

27. J. J. M. van Griethuysen et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.* **77**(21), e104–e107 (2017).
28. A. R. Cunliffe et al., "Lung texture in serial thoracic CT scans: assessment of change introduced by image registration," *Med. Phys.* **39**(8), 4679–4690 (2012).
29. R. M. Haralick, S. Shanmugam, and I. Sinstein, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern.* **SMC-3**(6), 610–621 (1973).
30. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, Vol. **1**, p. 460, Addison-Wesley Longman Publishing Co., Inc., Boston (1992).
31. G. Rodrigues et al., "Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review," *Radiother. Oncol.* **71**(2), 127–138 (2004).
32. E. D. York et al., "Dose–volume factors contributing to the incidence of radiation pneumonitis in non-small-cell lung cancer patients treated with three-dimensional conformal radiation therapy," *Int. J. Radiat. Oncol. Biol. Phys.* **54**(2), 329–339 (2002).
33. H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control* **AC-19**, 716–723 (1974).
34. Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypothesis," *Econometrica* **57**(2), 307–333 (1989).
35. M. Vallières et al., "Responsible radiomics research for faster clinical translation," *J. Nucl. Med.* **59**(2), 189–193 (2018).

Joseph J. Foy is a PhD candidate at The University of Chicago studying in the Graduate Program in Medical Physics. His research focuses on the variability in the radiomics and texture analysis workflow, and he is working to help standardize the radiomics research across institutions to allow for greater clinical implementation. In particular, his research has focused on the variation in radiomics features due to differences in radiomics software packages, image acquisition parameters, and reconstruction methods.

Samuel G. Armato III is an associate professor of radiology and a member of the Committee on Medical Physics at The University of Chicago. His research interests involve the development of computer-aided diagnostic (CAD) methods for thoracic imaging, including automated lung nodule detection and analysis in CT scans, semiautomated mesothelioma tumor response assessment, image-based techniques for the assessment of radiotherapy-induced normal tissue complications, and the automated detection of pathologic change in temporal subtraction images.

Hania A. Al-Hallaq is an associate professor of radiation oncology at The University of Chicago. She investigates the use of medical images to: 1) inform treatment selection, 2) guide treatment positioning, and 3) assess treatment response following radiotherapy. She combines her research background in texture and image analysis with her experience as a clinical radiotherapy physicist to contribute significantly to translational cancer research, including the study of radiomics to assess normal tissue toxicities following radiotherapy.