Assessment of Renal Cell Carcinoma by Texture Analysis in Clinical Practice: A Six-Site, Six-Platform Analysis of Reliability

Ankur M. Doshi, MD[1], Angela Tong, MD[1], Matthew S. Davenport, MD[2], Ahmed M. Khalaf, MD[3], Rafah Mresh, MD[3], Henry Rusinek, PhD[1], Nicola Schieda, MD[4], Atul B. Shinagare, MD[5], Andrew D. Smith, MD, PhD[3], Rebecca Thornhill, PhD[4], Raghunandan Vikram, MD[6], Hersh Chandarana, MD[1]

Genitourinary Imaging · Original Research

Keywords

CT, radiomics, renal cell carcinoma, texture analysis

Submitted: Jan 5, 2021
Revision requested: Feb 1, 2021
Revision received: Mar 10, 2021
Accepted: Apr 2, 2021
First published online: Apr 14, 2021

H. Rusinek is a codeveloper of FireVoxel. The remaining authors declare that they have no disclosures relevant to the subject matter of this article.

Based on a presentation at the Society of Abdominal Radiology 2019 annual meeting, Orlando, FL.

BACKGROUND. Multiple commercial and open-source software applications are available for texture analysis. Nonstandard techniques can cause undesirable variability that impedes result reproducibility and limits clinical utility.

OBJECTIVE. The purpose of this study is to measure agreement of texture metrics extracted by six software packages.

METHODS. This retrospective study included 40 renal cell carcinomas with contrast-enhanced CT from The Cancer Genome Atlas and Imaging Archive. Images were analyzed by seven readers at six sites. Each reader used one of six software packages to extract commonly studied texture features. Inter- and intrareader agreement for segmentation was assessed with intraclass correlation coefficients (ICCs). First-order (available in six packages) and second-order (available in three packages) texture features were compared between software pairs using Pearson correlation.

RESULTS. Inter- and intrareader agreement was excellent (ICC, 0.93–1). First-order feature correlations were strong ($r \geq 0.8$, $p < .001$) between 75% (21/28) of software pairs for mean intensity and SD, 48% (10/21) for entropy, 29% (8/28) for skewness, and 25% (7/28) for kurtosis. Of 15 second-order features, only cooccurrence matrix correlation, gray-level nonuniformity, and run-length nonuniformity showed strong correlation between software packages ($r = 0.90–1$, $p < .001$).

CONCLUSION. Variability in first- and second-order texture features was common across software configurations and produced inconsistent results. Standardized algorithms and reporting methods are needed before texture data can be reliably used for clinical applications.

CLINICAL IMPACT. It is important to be aware of variability related to texture software processing and configuration when reporting and comparing outputs.

Quantitative imaging techniques are valuable for the diagnosis and management of disease [1]. Texture analysis is performed through extraction and statistical analysis of pixel intensity and position data. Several investigations have shown that texture features may be valuable for the analysis of renal masses, including differentiation of benign from malignant lesions [2–5], prediction of renal cell carcinoma grade [6–9], gene mutation [10], and response to therapy in metastatic renal cell carcinoma [11]. Texture analysis has also been used to evaluate pathology in other organ systems, including the liver, pancreas, bowel, and lung [12].

Potential sources of variation in texture analysis include image acquisition parameters, segmentation methods, image postprocessing algorithms, and methods of statistical analysis [13–15]. A variety of texture software packages are available that apply imaging filters to remove noise, enhance edges, standardize gray-levels, and allow selection of fine, medium, or coarse texture features [12]. Software packages have variable control of settings, algorithms, and quality of available documentation, making it difficult to compare texture outputs across studies [12].

First-order texture features are determined according to histogram analysis of pixel intensities, whereas second-order texture features involve statistical analysis of patterns of

[1]Department of Radiology, New York University Langone Medical Center, 660 First Ave, New York, NY 10016. Address correspondence to A. M. Doshi (ankur.doshi@nyulangone.org).

[2]Department of Radiology, University of Michigan Health Systems, Ann Arbor, MI.

[3]Department of Radiology, University of Alabama at Birmingham, Birmingham, AL.

[4]Department of Medical Imaging, The Ottawa Hospital, The University of Ottawa, Ottawa, ON, Canada.

[5]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

[6]Department of Diagnostic Radiology, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

doi.org/10.2214/AJR.21.25456
AJR 2021; 217:1132–1140
ISSN-L 0361–803X/21/2175–1132

spatial distribution of intensity levels in a neighborhood [12]. Variability in computations and lack of standardization may impede the reproducibility of outputs across software platforms. Historical lack of standards for texture analysis is a recognized challenge for clinical implementation [16]. Outputs must be evaluated to ensure they can serve as standardizable biomarkers usable in clinical practice or clinical trials [17].

The purpose of this investigation was to measure the agreement of texture features extracted from six commercial and open-source software packages using the same CT dataset of clear cell renal cell carcinoma (ccRCC).

Methods

Patient Selection

Institutional review board approval was not required because retrospective data were retrieved from The Kidney Renal Clear Cell Carcinoma database, a publicly available dataset of de-identified images and clinical data from The Cancer Genome Atlas and The Cancer Imaging Archive [18, 19]. Contrast-enhanced CT scans in the nephrographic phase with 2.5-mm slices were included. Twenty patients with low-grade ccRCC (Fuhrman grade 1 or 2) and 20 patients with high-grade ccRCC (Fuhrman grade 3 or 4) were randomly selected for inclusion in this study. Patient demographics and tumor characteristics are provided in Table 1.

Multiplatform, Multisite Analysis

This study was conducted by members of the Society of Abdominal Radiology Disease Focused Panel on Renal Cell Carcinoma. A total of six texture software packages were used in this study, including FireVoxel (Build 339), TexRAD (version 3.9.2867.1553, Feedback Medical), Liver Fat Quantification (LFQ version 1.0, Liver Nodularity), ImageJ (version 1.52h, National Institutes of Health) with a locally developed plug-in to extract entropy [20], MaZda (version 4.6, Institute of Electronics, Technical University of Lodz), and 3D Slicer (version 4.8.1) with a radiomics extension based on the Pyradiomics library, which will be referred to as 3D Slicer [21, 22]. Sites were assigned the software according to preexisting availability, with the exception of 3D Slicer, which was specifically downloaded for a site without an existing texture analysis software.

HIGHLIGHTS

Key Finding

- *Six texture software packages using the same CT data produced variable results for first-order features, with strong correlations in 75% of software pairs for mean intensity and SD, 48% for entropy, 29% for skewness, and 25% for kurtosis. Three of 15 second-order texture features showed strong correlations.*

Importance

- *Variability in texture software outputs highlights the importance of efforts to standardize approaches to analysis and using caution when comparing results from different software.*

Image Preparation

An abdominal imaging fellow (A.T.) used FireVoxel to manually perform a 2D segmentation of each renal mass on a single axial CT slice on which the mass was largest (Fig. 1). Screenshots of masses with ROIs and CT DICOM files were distributed to each site. Because of intersoftware incompatibility regarding ROI file formats, it was not possible to export segmentation masks and import them into all remaining software, which would have eliminated any variability related to manual segmentation.

Image Segmentation

Inter- and intrareader agreement for the tumor segmentation process was assessed to confirm that segmentation is not a significant source of variability. LFQ was used by two independent readers (A.M.K. and R.M.) and TexRAD was used by two independent readers (A.B.S. and R.V.) to evaluate for interreader agreement. FireVoxel was used by the same reader (A.T.) in two separate sessions to assess for intrareader agreement.

Texture Analysis

Four of the readers were board-certified radiologists (M.S.D., N.S., A.B.S., and R.V.), one was an abdominal imaging fellow (A.T.),

TABLE 1: Patient Demographics and Tumor Characteristics

Characteristic	Fuhrman Grade of ccRCC		Total (*n* = 40)
	1–2 (*n* = 20)	3–4 (*n* = 20)	
Age (y), mean ± SD	56 ± 6.5	58 ± 13.9	57 ± 10.7
Sex			
Male	11 (55)	14 (70)	25 (62)
Female	9 (45)	6 (30)	15 (38)
Location of mass			
Right side	7 (35)	14 (70)	21 (52)
Left side	13 (65)	6 (30)	19 (48)
Diameter (cm), mean ± SD	5.0 ± 2.5	7.0 ± 2.6	6.0 ± 2.8

Note—Unless otherwise indicated, values are number (%). ccRCC = clear cell renal cell carcinoma.

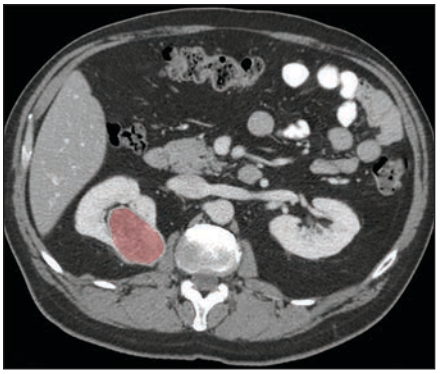

Fig. 1—Screen shot of ROI (*red area*) over mass in right kidney on single axial CT slice. ROI was manually drawn, avoiding edge on CT slice on which tumor was largest.

two were medical researchers (A.M.K. and R.M.) and one was a doctoral research scientist (R.T). Details regarding the software settings to apply were distributed to all the readers by the study leader (A.M.D.), who is a board-certified radiologist who has conducted prior research in texture analysis. Readers performed the following tasks using their assigned texture package for each mass: loaded the provided DICOM slice into the texture software, used the provided screen shot as a guide to perform a 2D segmentation of the mass and avoid the edges, performed texture analysis without application of any additional image filters if provided with the option (recognizing that filters could potentially be applied automatically without user control), set a distance factor of 1 pixel for software capable of second-order metrics, and sent the texture metric output files back to the data manager. For the site using MaZda and ImageJ, segmentations were performed in ImageJ and feature extraction was performed in MaZda using a previously established pipeline that has been used for evaluation of renal masses [5, 23]. For 3D Slicer and FireVoxel, bin width value was set to 5 and the option to enforce symmetric gray-level cooccurrence matrix (GLCM) was checked. For TexRAD, the Solid-Lesion_CT algorithm and polygon ROI settings were used. MaZda normalization was set to default. For ImageJ, the automatic binning option was selected and the maximum number of bins was set at 256. Otherwise, default software settings were used and no additional image normalization techniques were applied.

Data Review and Normalization

The output metrics from each software were reviewed to identify common metrics that could be compared across programs. An initial review of the output data from each software package was performed to determine whether normalization was required. If there was a consistent trend of data variability among the output metrics from each software, imaging DICOM data and available software manuals were reviewed to identify any embedded data conversions used in the calculations. Normalization consisted of applying an offset or a scaling factor when applicable (see Results for details). If normalization was not applicable, comparison of results included all outliers. Normalized and uncorrected first-order data are provided in Table 2.

Software Agreement

For each mass and each texture feature, the median value was computed. A relative percentage difference from the median was computed as 100 times the absolute value of V minus V_m divided by V_m (100 $|V - V_m| / V_m$), where V_m was the median of all software outputs and V was the output from an individual software pro-

TABLE 2: Mean Intensity and SD of 40 Renal Masses Stratified by Texture Software Package With Uncorrected and Normalized Values

First-Order Feature	FireVoxel	TexRAD		LFQ		ImageJ	MaZda	3D Slicer
		Reader 1	Reader 2	Reader 1	Reader 2			
Mean intensity (HU)								
Uncorrected	256.01 (366.08)	257.18 (366.32)	257.18 (365.62)	256.07 (366.09)	256.40 (366.04)	256.51 (365.62)	171.28 (20.37)	257.02 (365.99)
Normalized	102.41 (25.66)	103.58 (24.66)	103.58 (24.61)	103.22 (25.66)	101.84 (25.99)	103.03 (25.88)	171.28 (20.37)	103.42 (25.39)
SD (uncorrected)	38.76 (8.65)	40.37 (7.44)	41.51 (6.92)	38.85 (8.79)	38.61 (8.64)	38.43 (8.40)	21.82 (7.43)	39.88 (7.89)
Entropy (uncorrected)	4.06 (0.19)	4.89 (0.34)	4.91 (0.33)	6.97 (0.54)	6.97 (0.53)	6.85 (0.49)	No data	4.94 (0.31)
Kurtosis								
Uncorrected	0.62 (3.73)	0.98 (3.52)	1.36 (3.85)	3.29×10^{-6} (6.96×10^{-6})	3.32×10^{-6} (6.88×10^{-6})	−0.14 (0.55)	0.05 (0.57)	3.77 (3.17)
Normalized 3D Slicer and rescaled LFQ	0.62 (3.73)	0.98 (3.52)	1.36 (3.85)	0.33 (0.70)	0.33 (0.70)	−0.14 (0.55)	0.05 (0.57)	0.77 (3.17)
Skewness								
Uncorrected	−0.11 (0.58)	−0.20 (0.58)	−0.24 (0.66)	-4.80×10^{-6} (2.10×10^{-6})	-4.86×10^{-6} (2.06×10^{-6})	0.60 (3.69)	−0.16 (0.41)	−0.16 (0.53)
Rescaled LFQ	−0.11 (0.58)	−0.20 (0.58)	−0.24 (0.66)	−0.48 (2.12)	−0.49 (2.09)	0.60 (3.69)	−0.16 (0.41)	−0.16 (0.53)

Note—Numbers are means with SD in parentheses. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with a locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with a radiomics extension based on the Pyradiomics library.

gram. There were instances in which skewness or kurtosis for a tumor had both negative and positive values depending on the software. To maintain positive values for the magnitude of the difference ratio, the absolute value was used. Mean percentage difference of all masses was computed for each texture measure and software program.

Intersoftware agreement of first-order texture features was also assessed in terms of relative agreement using the Pearson correlation between results produced by each pair of software packages. A total of 61 correlations were tested for each pair of software packages. As a result, correlations for each pair would remain significant after a Bonferroni multiple comparison correction only if associated with a p value less than .05 / 61, or approximately 0.00082. Results were considered significant at the Bonferroni corrected significance level if the p value was less than .001. The Pearson rank correlation was used to compare second-order texture metric outputs from FireVoxel, MaZda, and 3D Slicer. Correlations, when positive, were considered strong ($r \geq 0.8$), moderate ($r \geq 0.5$ and < 0.8), weak ($r \geq 0.2$ and < 0.5), or uncorrelated ($r < 0.2$) [24].

Inter- and Intrareader Agreement

For the two readers using TexRAD, two readers using LFQ, and single reader using FireVoxel twice, interreader and intrareader agreement were calculated by intraclass correlation coefficient (ICC) for absolute (literal) agreement among single measures. ICC values were considered to reflect poor, moderate, good, or excellent reliability according to values of less than 0.5, 0.5–0.74, 0.75–0.9, and greater than 0.9, respectively [25].

Statistical tests were conducted using SAS 9.4 software (SAS Institute) and MedCalc (version 19.5.1). Two-sided 5% significance levels were used for tests without Bonferroni correction. Correlation matrix figures were generated using Displayr (Displayr).

Results

Data Inventory

Mean intensity, SD, skewness, and kurtosis were available for all software programs. Entropy was available for all except MaZda. Second-order texture features that were precisely documented for MaZda, FireVoxel, and 3D Slicer were included ($n = 15$): joint energy (angular second moment), contrast, cooccurrence matrix correlation, sum of squares, inverse difference moment, sum average, sum entropy, joint entropy, difference variance, difference entropy, gray-level nonuniformity, run percentage (average fraction), long-run emphasis, run-length nonuniformity, and short-run emphasis.

Segmentation Interreader and Intrareader Agreement

There was excellent agreement between the two independent readers using TexRAD (ICC, 0.93–0.99), two independent readers using LFQ (ICC, 0.99 for all measures), and single reader using FireVoxel twice (ICC, 0.98–1).

First-Order Texture Metric Comparison

Mean intensity—Initial review of the mean intensity showed that values were approximately 1024 HU greater for the same six masses in TexRAD, LFQ, ImageJ, FireVoxel, and 3D Slicer. This was likely because of a rescale intercept setting in a DICOM tag. For these instances, the data were normalized by subtracting 1024 HU from the reported mean value. The relative percentage difference from the median was 71.59% for MaZda but was 1.09–2.14% for the remaining packages (Table 3). The Pearson correlation matrix of means produced by each pair of software packages is shown in Figure 2. Correlations were strong for 75% (21/28) of software pairs and moderate for 25% (7/28). FireVoxel, ImageJ, LFQ, 3D Slicer, and TexRAD showed strong correlations ($r = 0.96$–1). Correlations were moderate between MaZda and all the other packages ($r = 0.72$–0.77). These correlations were all statistically significant ($p < .001$).

SD—The MaZda platform reported variance, which was converted to SD by taking the square root. SD values from MaZda were outliers, and no data normalization techniques were applied. Percentage difference from the median for MaZda was 43.46%. The remaining programs showed differences of 1.14–9.03% (Table 3). Pearson correlations were strong and significant ($r = 0.87$–1, $p < .001$) for 75% (21/28) of software pairs, but weak

TABLE 3: Relative Percentage Difference From Cross-Platform Median Stratified by Texture Software Package

Platform	Mean Intensity	SD	Entropy	Kurtosis	Skewness
FireVoxel	1.36	1.14	18.77	428.08	46.28
TexRAD					
Reader 1	1.69	5.05	2.50	298.73	60.01
Reader 2	2.14	9.03	2.06	342.85	79.13
LFQ					
Reader 1	2.02	1.14	39.03	115.31	136.53
Reader 2	2.13	1.21	38.98	117.74	119.52
ImageJ	1.58	2.04	36.69	504.76	408.00
MaZda	71.59	43.46	No data	450.07	25.22
3D Slicer	1.09	3.51	1.36	289.73	49.25

Note—FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with a locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with a radiomics extension based on the Pyradiomics library.

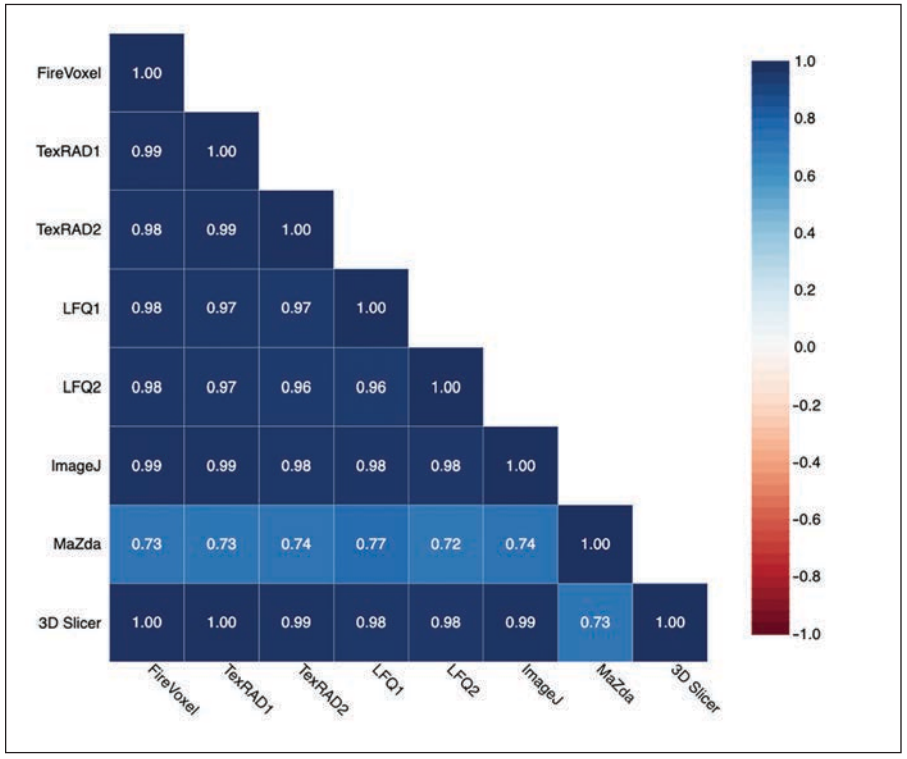

Fig. 2—Pearson correlation matrix of mean intensity produced by each pair of software packages. Numbers after name of package indicate reader 1 or 2. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with radiomics extension based on Pyradiomics library.

and nonsignificant after Bonferroni correction (r = 0.43–0.49, p = .001–.005) for 25% (7/28) (Fig. 3).

Entropy—MaZda did not provide entropy data. Percent differences from the median ranged from 18.77 to 39.03% for FireVoxel, LFQ, and ImageJ and ranged from 1.36 to 2.50% for 3D Slicer and TexRAD (Table 3). Pearson correlations were strong and significant (r = 0.84–1, p < .001) for 48% (10/21) of software pairs and moderate and significant (r = 0.62–0.78, p < .001) for 29% (6/21) (Fig. 4). Correlations were weak and nonsignificant (r = 0.28–0.29, p = .07–.09) for 24% (5/21) of software pairs.

Kurtosis—Initial data review showed that kurtosis values from 3D Slicer were approximately 3 units greater than data from other packages. Documentation for the 3D Slicer radiomics extension indicates that the Image Biomarker Standardization Ini-

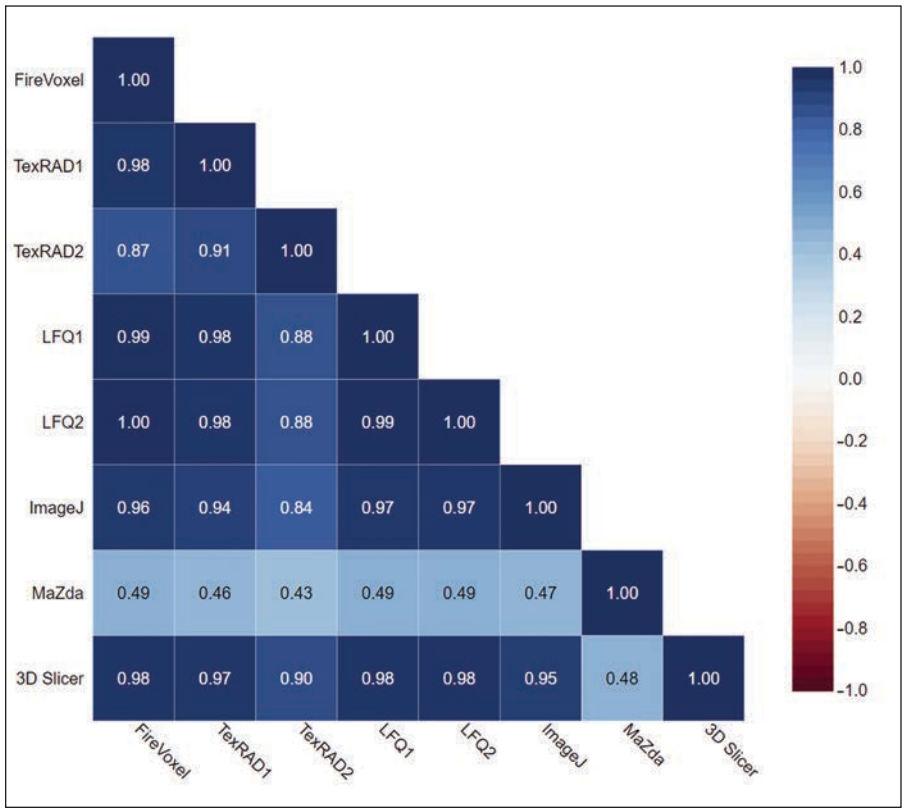

Fig. 3—Pearson correlation matrix of SD produced by each pair of software packages. Numbers after name of package indicate reader 1 or 2. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with radiomics extension based on Pyradiomics library.

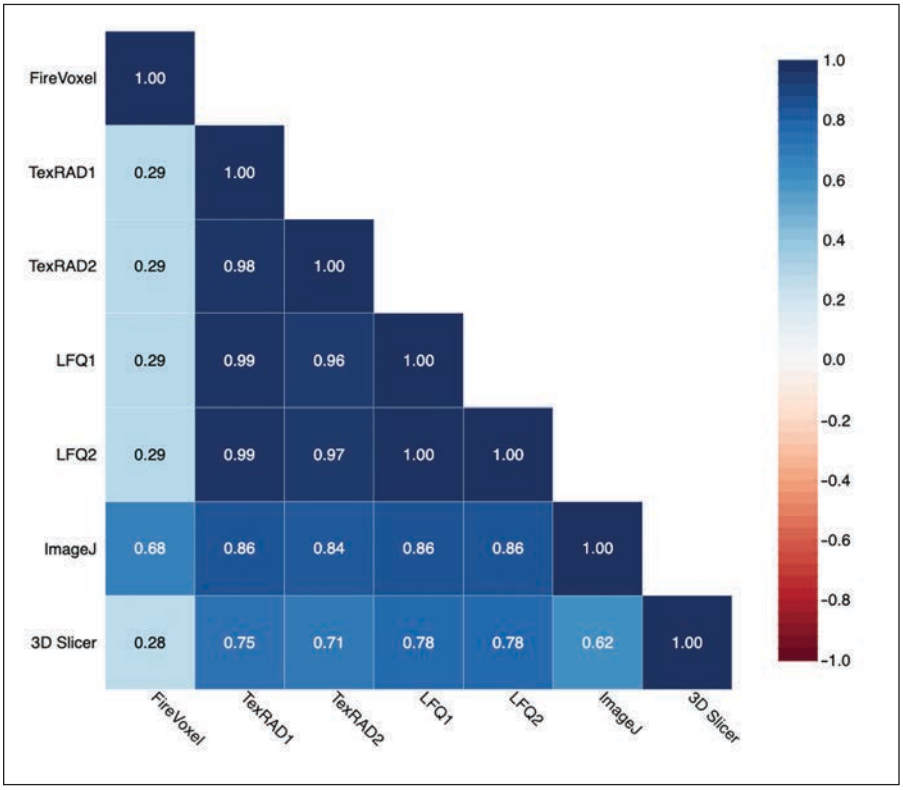

Fig. 4—Pearson correlation matrix of entropy produced by each pair of software packages. Numbers after name of package indicate reader 1 or 2. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with locally developed plug-in to extract entropy; 3D Slicer = 3D Slicer version 4.8.1 with radiomics extension based on Pyradiomics library.

tiative (IBSI) definition of kurtosis involves subtracting by 3 to center on 0 for normal distributions, whereas with Pyradiomics kurtosis is not corrected [26]. Therefore, results from 3D Slicer were corrected by subtracting 3.

Kurtosis values from LFQ were on the order of 10^{-4}, and normalization was achieved by rescaling by multiplying each value by 10^4. Percent differences from the median for all software packages ranged from 115.31–504.76% (Table 3). Correlations were strong and significant (r = 0.96–1, p < .001) for 25% (7/28) of software pairs, and moderate and significant (r = 0.64–0.72, p < .001) for 21% (6/28). Correlations were weak (r = 0.20–0.45) for 26% (10/28) of software pairs and uncorrelated (r = −0.16 to 0.13) for 18% (5/28); these relationships were not significant after Bonferroni correction (p = .003–.74) (Fig. 5).

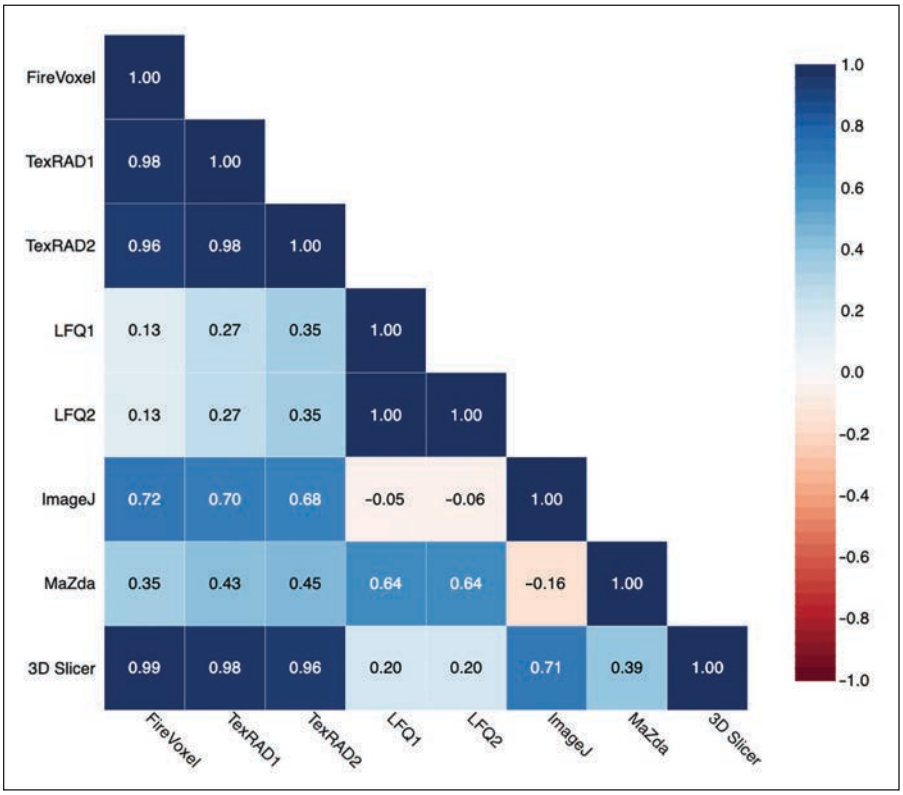

Fig. 5—Pearson correlation matrix of kurtosis produced by each pair of software packages. Numbers after name of package indicate reader 1 or 2. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with radiomics extension based on Pyradiomics library.

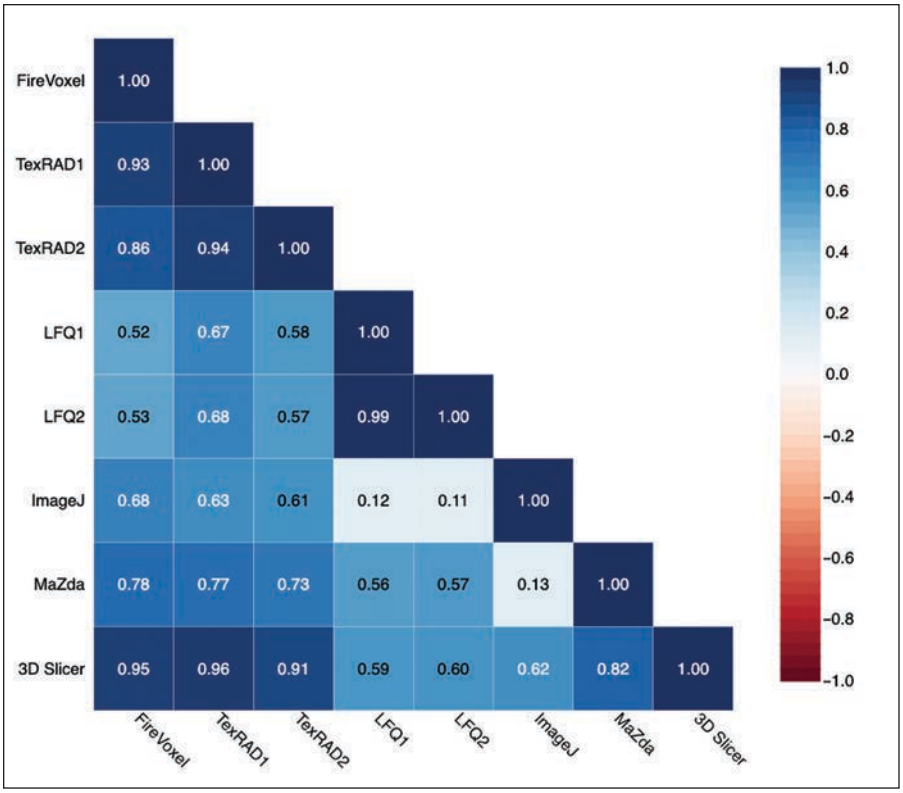

Fig. 6—Pearson correlation matrix of skewness produced by each pair of software packages. Numbers after name of package indicate reader 1 or 2. FireVoxel = FireVoxel Build 339; TexRAD = TexRAD version 3.9.2867.1553, Feedback Medical; LFQ = Liver Fat Quantification version 1.0, Liver Nodularity; ImageJ = ImageJ version 1.52h, National Institutes of Health with locally developed plug-in to extract entropy; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with radiomics extension based on Pyradiomics library.

Skewness—Initial data review showed that skewness values for LFQ were on the order of 10^{-4} and therefore were rescaled by multiplying each value by 10^4. Percentage differences ranged from 25.22% to 408.00% (Table 3). Correlations were strong and significant (r = 0.82–1, $p < .001$) for 29% (8/28) of software pairs and moderate (r = 0.52–0.78, $p < .001$) for 61% (17/28). Results were uncorrelated (r = 0.11–0.13, p = .49–.54) for 11% (3/28) (Fig. 6).

Second-Order Texture Features

Comparison of mean outputs produced by the packages are provided in Table 4, and results of correlation tests for second-order texture features are listed in Table 5. Most (12/15) features showed very strong ($r > 0.95$, $p < .001$) correlations between 3D Slicer and FireVoxel. Most results from these two packages showed weak or no correlation to MaZda results. Only cooccurrence matrix correlation, gray-level nonuniformity, and run-length nonuniformity showed strong correlation between all three packages (r = 0.90–1.00; $p < .001$).

Discussion

Our multicenter multiplatform study showed that texture features are inconsistent across software platforms. This is problematic because quantitative measures must be reliable and standardized to be considered clinically relevant. Simple features with unambiguous definitions (e.g., mean intensity) showed unexpected variability. Mean intensity was approximately 1024 HU greater in some masses, likely related to a rescale intercept setting in the DICOM data. Mean intensity reported by MaZda, which was initially developed for MRI analysis, was not highly correlated to other packages, possibly because of a default normalization step, inability to import negative pixel values, or lack of utilization of DICOM tag information [27].

Although all software packages reported first-order kurtosis, 3D Slicer used a computation that did not implement excess kurtosis and required a correction. This variability highlights the importance of exploring differences of feature nomenclature or nonconventional computations. A minority (3/15) of second-order features showed strong correlations between FireVoxel, 3D Slicer, and MaZda. Liang et al. [28] found that first-order and GLCM features showed greater correlation than gray-level run-length matrix and gray-level size-zone matrix features. The authors suggested the greater reliability of first-order features is because higher-order textural features are sensitive to preprocessing steps.

Potential sources of variability encountered during our cross-platform analysis included image filtration, data scaling, data normalization, histogram binning, and unique mathematic algorithms. Software platforms may offer variable user control of these parameters. Gray-level intensities can be organized into discrete bins by choosing either a fixed number or size of bins, and texture values can be impacted by the method used [29–31]. Schwier et al. [32] found that repeatability of radiomic features of prostate tumors on MRI is sensitive to processing parameters. Foy et al. [27] compared outputs from four software packages and found that first- and second-order features showed differences because of variations in image importation and preprocessing, algorithm implementation, feature naming conventions, and gray-level cooccurrence matrix parameters.

There is a recognized need for standardization in texture analysis before it can be applied in clinical practice [12]. Investi-

TABLE 4: Mean (SD) of Second-Order Texture Features and the Test for Equality of the Means

Feature	Mean (SD)			3D Slicer – Mazda		MaZda–FireVoxel		3D Slicer–FireVoxel	
	3D Slicer	Mazda	FireVoxel	Mean	p^a	Mean	p^a	Mean	p^a
GLCM contrast	50.32 (26.14)	6.35 (4.15)	11.9 (6.3)	43.96	< .001	–5.51	.26	38.44	< .001
GLCM correlation	0.59 (0.18)	0.58 (0.20)	0.58 (0.19)	0.01	.91	–0.00	.99	0.01	.95
Difference entropy	3.80 (0.34)	0.71 (0.15)	2.83 (0.33)	3.09	< .001	–2.11	< .001	0.97	< .001
Difference variance	19.02 (9.80)	2.45 (1.49)	4.45 (2.23)	16.57	< .001	–2.00	.28	14.57	< .001
Gray-level nonuniformity	134.2 (116.1)	319.3 (304.5)	222.0 (199.0)	–185.1	< .001	97.4	.12	–87.8	.18
Inverse difference moment	0.18 (0.04)	0.43 (0.13)	0.32 (0.07)	–0.24	< .001	0.11	< .001	–0.14	< .001
Joint energy	0.0027 (0.0018)	0.0300 (0.0442)	0.0098 (0.0055)	–0.0271	< .001	0.0201	.02	–0.0071	.43
Joint entropy	8.93 (1.57)	1.87 (0.36)	7.26 (0.62)	7.06	< .001	–5.40	< .001	1.66	< .001
Long-run emphasis	1.22 (0.07)	2.34 (2.07)	1.51 (0.19)	–1.12	< .001	0.83	.007	–0.29	.53
Run-length nonuniformity	3263.4 (2918.8)	1951.4 (1972.4)	2379.6 (2222.1)	1312.0	.04	–428.2	.70	883.7	.23
Run percentage	0.94 (0.02)	0.80 (0.11)	0.87 (0.04)	0.13	< .001	–0.08	< .001	0.06	< .001
Short-run emphasis	0.95 (0.01)	0.84 (0.09)	0.90 (0.03)	0.11	< .001	–0.06	< .001	0.05	< .001
Sum average	62.39 (16.18)	43.67 (5.10)	41.36 (5.03)	18.73	< .001	2.31	.57	21.03	< .001
Sum entropy	5.72 (0.34)	1.24 (0.20)	4.66 (0.35)	4.48	< .001	–3.43	< .001	1.06	< .001
Sum of squares	65.16 (25.23)	8.37 (4.52)	15.11 (6.22)	56.79	< .001	–6.75	.12	50.04	< .001

Note—3D Slicer = 3D Slicer version 4.8.1 with a radiomics extension based on the Pyradiomics library; MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; FireVoxel = FireVoxel Build 339. GLCM = gray-level cooccurrence matrix.

[a]Adjusted *p* values were calculated with analysis performed using three-group ANOVA followed by Tukey range post hoc test to identify comparisons of means that are significantly different.

TABLE 5: Pearson Correlation (*r*) and Significance Level (*p*) Between MaZda, 3D Slicer, and FireVoxel for Second-Order Texture Metrics

Measure	MaZda vs 3D Slicer		3D Slicer vs FireVoxel		FireVoxel vs MaZda	
	r	*p*	*r*	*p*	*r*	*p*
Contrast	0.69	< .001	0.99	< .001	0.68	< .001
Cooccurrence matrix correlation	0.97	< .001	0.98	< .001	0.98	< .001
Difference entropy	0.45	.004	0.99	< .001	0.43	.005
Difference variance	0.67	< .001	0.97	< .001	0.66	< .001
Gray-level nonuniformity	0.90	< .001	1.00	< .001	0.91	< .001
Inverse difference moment	0.41	.009	0.99	< .001	0.39	.01
Joint energy	–0.06	.71	0.78	.14	0.17	.31
Joint entropy	–0.03	.84	0.24	< .001	0.28	.08
Long-run emphasis	0.18	.25	0.98	< .001	0.19	.23
Run-length nonuniformity	0.93	< .001	0.99	< .001	0.92	< .001
Run percentage	0.29	.07	0.98	< .001	0.30	.06
Short-run emphasis	0.29	.07	0.98	< .001	0.30	.06
Sum average	–0.03	.87	0.08	.62	0.74	< .001
Sum entropy	0.41	.008	0.98	< .001	0.39	.01
Sum of squares	0.52	.001	0.96	< .001	0.53	< .001

Note—MaZda = MaZda version 4.6, Institute of Electronics, Technical University of Lodz; 3D Slicer = 3D Slicer version 4.8.1 with a radiomics extension based on the Pyradiomics library; FireVoxel = FireVoxel Build 339.

gators must exercise caution when comparing numeric values across software packages unless processing details and configurations are known. Lack of reproducibility and validation of quantitative imaging studies is a recognized limitation of radiomics. The Image Biomarker Standardization Initiative (IBSI) seeks to establish standards including nomenclature, benchmarks, and reporting guidelines [33].

Our study had several limitations. Intrareader agreement was tested for one reader and on one platform. Although manual segmentation can be a potential source of variation [34], our study

showed excellent inter- and intrareader agreement for the tested readers and platforms, indicating that manual segmentation was not a significant component of variation. Detailed manuals or source codes were not available for all software packages, making it challenging to explain differences in some of the measures. Some software packages did not allow user control of certain settings, making it difficult to standardize the processing. In particular, a variety of gray-level discretization methods were provided, which could impact consistency of certain measures, especially entropy and second-order features.

In conclusion, our results show variation in how software packages process image data and compute texture features. Investigators must exercise caution when comparing results from studies using different software packages. Future radiomic investigations should report image processing steps, software settings, and any nonconventional computations used.

References

1. Rosenkrantz AB, Mendiratta-Lala M, Bartholmai BJ, et al. Clinical utility of quantitative imaging. *Acad Radiol* 2015; 22:33–49
2. Raman SP, Chen Y, Schroeder JL, Huang P, Fishman EK. CT texture analysis of renal masses: pilot study using random forest classification for prediction of pathology. *Acad Radiol* 2014; 21:1587–1596
3. Erdim C, Yardimci AH, Bektas CT, et al. Prediction of benign and malignant solid renal masses: machine learning-based CT texture analysis. *Acad Radiol* 2020; 27:1422–1429
4. Feng Z, Rong P, Cao P, et al. Machine learning-based quantitative texture analysis of CT images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *Eur Radiol* 2018; 28:1625–1633
5. Hodgdon T, McInnes MD, Schieda N, Flood TA, Lamb L, Thornhill RE. Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images? *Radiology* 2015; 276:787–796
6. Deng Y, Soule E, Samuel A, et al. CT texture analysis in the differentiation of major renal cell carcinoma subtypes and correlation with Fuhrman grade. *Eur Radiol* 2019; 29:6922–6929
7. Ding J, Xing Z, Jiang Z, et al. CT-based radiomic model predicts high grade of clear cell renal cell carcinoma. *Eur J Radiol* 2018; 103:51–56
8. Kocak B, Durmaz ES, Ates E, Kaya OK, Kilickesmez O. Unenhanced CT texture analysis of clear cell renal cell carcinomas: a machine learning-based study for predicting histopathologic nuclear grade. *AJR* 2019; 212:[web] W132–W139
9. Feng Z, Shen Q, Li Y, Hu Z. CT texture analysis: a potential tool for predicting the Fuhrman grade of clear-cell renal carcinoma. *Cancer Imaging* 2019; 19:6
10. Kocak B, Durmaz ES, Ates E, Ulusan MB. Radiogenomics in clear cell renal cell carcinoma: machine learning-based high-dimensional quantitative CT texture analysis in predicting PBRM1 mutation status. *AJR* 2019; 212:[web]W55–W63
11. Goh V, Ganeshan B, Nathan P, Juttla JK, Vinayan A, Miles KA. Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker. *Radiology* 2011; 261:165–171
12. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT texture analysis: definitions, applications, biologic correlates, and challenges. *RadioGraphics* 2017; 37:1483–1503
13. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys* 2018; 102:1143–1158
14. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology* 2018; 288:407–415
15. Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Texture analysis of imaging: what radiologists need to know. *AJR* 2019; 212:520–528
16. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2016; 278:563–577
17. Heye T, Davenport MS, Horvath JJ, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part I. Perfusion characteristics in the female pelvis by using multiple computer-aided diagnosis perfusion analysis solutions. *Radiology* 2013; 266:801–811
18. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013; 26:1045–1057
19. Kirby J, Jarosz Q. TCGA-KIRC. The Cancer Imaging Archive website. wiki.cancerimagingarchive.net/display/Public/TCGA-KIRC. Published 2016. Accessed December 6, 2017
20. Schieda N, Krishna S, McInnes MDF, et al. Utility of MRI to differentiate clear cell renal cell carcinoma adrenal metastases from adrenal adenomas. *AJR* 2017; 209:[web]W152–W159
21. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 2012; 30:1323–1341
22. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017; 77:e104–e107
23. Schieda N, Lim RS, Krishna S, McInnes MDF, Flood TA, Thornhill RE. Diagnostic accuracy of unenhanced CT analysis to differentiate low-grade from high-grade chromophobe renal cell carcinoma. *AJR* 2018; 210:1079–1087
24. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. *Radiology* 2003; 227:617–622
25. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15:155–163
26. Pyradiomics website. Radiomic features. pyradiomics.readthedocs.io/en/latest/features.html. Accessed March 31, 2021
27. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG 3rd. Variation in algorithm implementation across radiomics software. *J Med Imaging (Bellingham)* 2018; 5:044505
28. Liang ZG, Tan HQ, Zhang F, et al. Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma. *Br J Radiol* 2019; 92:20190271
29. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol* 2016; 61:R150–R166
30. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. *Eur Radiol* 2015; 25:2805–2812
31. Desseroit MC, Tixier F, Weber WA, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. *J Nucl Med* 2017; 58:406–411
32. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of multiparametric prostate MRI radiomics features. *Sci Rep* 2019; 9:9441
33. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020; 295:328–338
34. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O. Reliability of single-slice-based 2D CT texture analysis of renal masses: influence of intra- and interobserver manual segmentation variability on radiomic feature reproducibility. *AJR* 2019; 213:377–383