OBJECTIVE. Radiologic texture is the variation in image intensities within an image and is an important part of radiomics. The objective of this article is to discuss some parameters that affect the performance of texture metrics and propose recommendations that can guide both the design and evaluation of future radiomics studies.

CONCLUSION. A variety of texture-extraction techniques are used to assess clinical imaging data. Currently, no consensus exists regarding workflow, including acquisition, extraction, or reporting of variable settings leading to poor reproducibility.

Texture Analysis

In material science, texture is defined as a measure of the variation of a surface; a rough-textured material would have a high rate of change in the high and low points of a surface, compared with a smooth-textured material [13]. In radiology, image texture refers to differences in the gray-scales representing an ROI. The image of a rough-textured material would have a high rate of change in the high and low points of a surface (the gray-scale value), compared with a smooth-textured material [14].

Typical radiomic assessment includes analysis of texture, shape, and size [1]. The underlying assumption of the technique is that the gray-scale values creating the image of the tumor and the spatial and temporal interrelationships of these values reflect the phenotypic variations of the tumor, which are indicative of genetic and other molecular variations [7].

Although there is a lot of interest in using radiomics for noninvasive tumor assessment, poor standardization and generalization of radiomic results hinder the translation of radiomics in clinical practice [8–12]. We conducted a PRISMA (preferred reporting items for systematic reviews and meta-analyses)–compliant search of the PubMed, Scopus, and Embase databases, using the following keywords: radiomic or radiomics (Fig. 1). No consensus currently exists regarding workflow, the reporting of parameter settings, or reproducibility [8–12]. In this article, we discuss some parameters that affect the performance of texture metrics. Our goal is to propose a set of recommendations that serve as guidelines to design future studies and help evaluate them.

Texture Analysis of Imaging: What Radiologists Need to Know

Bino A. Varghese1
Steven Y. Cen
Darryl H. Hwang
Vinay A. Duddalwar

Keywords: quantitative imaging, radiomics, texture

doi.org/10.2214/AJR.18.20624
Received September 6, 2018; accepted without revision September 6, 2018.

1All authors: Department of Radiology, University of Southern California, 1441 Eastlake Ave, Los Angeles, CA 90033. Address correspondence to V. A. Duddalwar (Vinay.Duddalwar@med.usc.edu).

AJR2019; 212:520–528
0361–803X/19/2123–520
© American Roentgen Ray Society
Texture Analysis of Imaging

However, when images are quantitatively assessed to extract meaningful data, variations in acquisition and image reconstruction parameters lead to inconsistent findings between different datasets, particularly in multicenter studies [18].

**Image Segmentation**

The image segmentation step involves identifying an ROI, which could be done automatically, semiautomatically, or manually. Although manual segmentation is accurate, it is more tedious and subjective [19]. Automatic segmentation is objective but error prone, especially when imaging artifacts and noise are encountered. Some of the commonly used automated segmentation algorithms include active contour–based [20], level set–based [21], and region- and graph-based methods [22, 23]. No established segmentation standard currently exists. More recently, deep learning techniques, such as convolutional...
neural networks, have been used for segmentation [24–26].

Texture analysis has been incorporated into the radiomics workflow at various stages. At the preprocessing stage, images could be segmented into contiguous regions on the basis of the texture properties of each region [27]; at the feature extraction and classification stages, texture features could provide cues for classifying patterns or identifying objects [8–11].

**Feature Extraction**

Statistical, transform-based, and structural-based texture assessment are the three main approaches used to describe texture (Table 1). Statistical characterization of texture is based on the assessment of texture as a measure of the statistical properties of the gray levels creating the ROI. These properties are conventionally computed from first-order statistical methods, such as histogram analysis in which the analysis is based on gray-scale values only and spatial information is lost. These first-order methods and analysis are relatively easier to implement and understand. Among the higher-order texture methods, which include both gray-scale values and spatial orientation, are the gray-level co-occurrence matrix and the gray-level difference matrix. In recent studies, gray-level run-length matrix metrics and gray-level size zone matrix metrics have also been reported [28]. Transform-based analysis involves extraction of texture metrics on the basis of properties of a wave spectrum and describes the global periodicity of the gray levels of a surface by identifying high-energy peaks in the spectrum and their variations. Structural methods involve techniques of decomposing an image into basic units and identifying the rules required to construct that given image from these basic units. Some examples of structural methods of texture assessment include fractal analysis [29].

The use of a large number of radiomic metrics and the lack of uniformity of these measures and their selective use, which may be correlated, have led to studies with results that are nonreproducible and noncomparable [8].

**Statistical Analysis**

The choice of statistical methods used in radiomics depends on multiple factors (e.g., whether the radiomic features are used as the outcome or the predictor or whether radiomics analysis is part of a pilot or confirmative study). When radiomic features are used as the outcome, the assumption of the data normality will need to be tested first. Statistical bias can be introduced on the basis of the choice of the statistical tests used as well as on the basis of the inherent noise and skewness of the medical data. Some of the commonly used statistical tests for normality include the *t* test, ANOVA, the Friedman test, the Kruskal-Wallis test, and the Mann-Whitney test [30]. If multiple groups or multiple features are involved, multiple testing errors should be controlled using techniques such as the Tukey honest significant difference test [31] or the Benjamini-Hochberg procedure [32]. A mixed-effect model should be used in case of longitudinal data. When radiomic features are used as the predictor, methods such as logistic regression analysis, ROC curve analysis, and the Cox regression model have been used in published radiomics studies for statistical analysis [33, 34]. When a pilot study with radiomic features is conducted, the large number of features usually overwhelms the sample size. If a multivariate analysis is performed, the rule of thumb is to include 10–25 subjects for each variable [35–37]. Some argue that there can be as few as two subjects; however, this modification could introduce bias for the model fitting, such as the $R^2$ model [38]. Variable selection tools, such as the least absolute shrinkage and selection operator, aid in enhancing the prediction accuracy and interpretability of the statistical model; however, they are not suitable for a pilot study with a small sample size because it is driven by the model-fitting index. Bujang et al. [39] suggested that to approximate the model $R^2$ from the true population, the minimum sample size needed for a multiple linear regression would be 300 subjects. This number is unlikely to be achieved in a pilot study. In such scenarios, univariate analysis of a large number of subjects may be a more preferred method. When a confirmation study with a large sample size is conducted, a large enough independent testing dataset will need to be reserved in advance. Then, on the basis of the size of the independent testing sample, a learning sample that is at least equal to or is 2–4 times the size of the testing sample must be obtained [40].

With the use of direction-based transformational texture metrics, such as Gabor filters, direction-dependent metrics can be correlated, leading to data redundancy. Unsupervised and supervised feature reduction methods could be used to reduce data redundancy. The unsupervised feature reduction methods include techniques such as principle component analysis, independent component analysis, zero variance, near-zero variance, and consensus clustering combined with principle component analysis [41]. When supervised machine learning is used, the feature reduction is implemented by selecting the variables with the best prediction value (i.e., the variable of importance).

**TABLE 1: Texture Metrics Extracted in the Radiomic Analysis of Imaging Data**

<table>
<thead>
<tr>
<th>Texture Metric</th>
<th>Method(s)</th>
<th>Descriptors</th>
<th>Imaging Modalities [Reference(s)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-order (statistical)</td>
<td>Histogram analysis</td>
<td>Mean, median, SD, kurtosis, skewness, quartiles, minimum, maximum, and other descriptors</td>
<td>CT [70, 83] MRI [84, 85] PET [86, 87] US [88, 89]</td>
</tr>
<tr>
<td>Transform analysis</td>
<td>Fourier, wavelets, discrete cosine, Gabor, and Law methods</td>
<td>Metrics assessing magnitude, phase, direction, and other descriptors</td>
<td>CT [83, 94] MRI [95] US [96]</td>
</tr>
<tr>
<td>Structural analysis</td>
<td>Fractal analysis</td>
<td>Fractal dimension</td>
<td>CT [87] MRI [98] PET [29] US [40]</td>
</tr>
</tbody>
</table>

*Note—US = ultrasound.*
A number of radiomics studies include scans from different centers. Although this increases the cohort size, it increases the number of variables and confounds the imaging data, leading to systematic errors and poor reliability. To this end, reliable metrics (i.e., metrics that are reproducible [i.e., their value remains unchanged across different scanners of a given imaging modality] and repeatable [i.e., their value remains the same when repeated multiple times on a single scanner]) need to be identified [42, 43] (Fig. 3). The two popular statistical indexes to assess reliability include the intraclass correlation coefficient (ICC) and the concordance correlation coefficient. When reproducibility alone is assessed without repeated measures for a given scanner or modality, the ICC2 (two-way random ICC) and ICC3 (two-way mixed ICC) are identical to the concordance correlation coefficient. However, if reproducibility is assessed with repeated measures, which is equivalent to assessing reproducibility and repeatability at once, only the ICC3 is identical to the concordance correlation coefficient [44]. In general, the concordance correlation coefficient or the ICC2 and ICC3 are the preferred assessment methods, with a heat map used as a visualization tool to illustrate the pattern of reliability.

Reliability Assessment of Texture Analysis Metrics

A number of researchers have studied the reliability of texture metrics derived from CT images and have concluded that radiomic features can be reproducible [5, 45–51]. Phantom studies assessing the interscanner and intrascanner differences of radiomic metrics have been reported using commercially available CT scanners and scanning protocols. Mackin et al. [45, 46] showed that the variability in radiomic metrics extracted from CT images of the Credence Cartridge Radiomics Phantom was comparable in size to the variability noted in the same features extracted from CT images of non–small cell lung carcinoma tumors. The interscanner variability of CT in radiomic metrics implies that the quality and repeatability of radiomic studies depends strongly on the consistency of image acquisition and reconstruction [45]. In a different study that used cone-beam CT, Fave et al. [48] showed that select radiomic metrics are robust to the noise and poor image quality of cone-beam CT images as long as consistent imaging protocols are used. This study also reports reasonable performance of the radiomics when limited breathing-related motion was involved [48]. A more recent study by Berenguer et al. [42] evaluated the reliability of CT-based texture metrics and reported that most radiomics metrics were redundant and nonreproducible. The investigators further summarized that if all the CT parameters are held constant, except for FOV, tube voltage, and tube current, then the information provided by the analyzed radiomic metrics could be summarized in 10 radiomic metrics only [42]. Although these studies are essential to advance the assessment of the reliability of radiomics as a clinical tool, they are limited to texture evaluation as opposed to comprehensive radiomic evaluation. Additional radiomic metrics such as size and shape, which have been shown to correlate with tumor behavior, have not been represented in these studies [49, 52, 53]. In addition, these phantom-study validations of the performance assessment of the reliable metrics have been conducted using single-phase unenhanced CT of patients with non–small cell lung carcinoma with fewer than 150 samples, which limits the scope and statistical power of the validation. The effect of using contrast enhancement dynamics on image texture has not been addressed.

In a tumor study, a database of tumor images acquired using different imaging protocols was used to investigate the repeatability and robustness of radiomic metrics in CT scans [49]. In that study, Zhao et al. [49] reported that good repeatability of features (with size, shape, and texture included) derived from non–small cell lung carcinoma scans with the same imaging settings; however, only 19% of 89 radiomic metrics were repeatable when six different imaging settings varying in slice thicknesses and reconstruction algorithms were used. The investigation was limited to a single CT scanner, and reproducibility of the radiomic metrics therefore was not assessed. In addition, the effects of changing CT acquisition parameters such as tube voltage, tube current, and pitch, which considerably affect image quality and thus potentially influence radiomic performance, also were not assessed.

Compared with the reliability of CT radiomics, the reliability of MRI radiomics has been assessed in a limited number of studies. A recent study by Yang et al. [54], performed to examine the dependence of image texture features on MRI acquisition parameters and reconstruction using a digital MRI phantom, showed that the coefficient of variation is substantial (> 20%) for some textural radiomic metrics, such as the gray-level difference matrix and the gray-level size zone matrix, but is comparatively lower for textural radiomic metrics, such as the gray-level co-occurrence matrix method. Results from the Yang et al. study indicate that not all radiomic metrics are equally susceptible to scanner- and imaging protocol–based changes. Some metrics may be reliable for combining multicenter MRI databases with images taken using different pulse sequences and MRI scanners to conduct radiomic studies with a large sam-

---

**Fig. 3**—Definition of reliable radiomic metrics. Radiomic metrics with low interscanner variation (reproducible metrics), low intrascanner variation (robust metrics), and high test-retest performance (repeatable metrics) form cohort of reliable radiomic metrics.
ple size. Mayerhoefer et al. [55] conducted a phantom study to assess the effect of MR acquisition parameters and found that gray-level co-occurrence matrix features were generally sensitive to variation in the mean number of scans obtained, TR, TE, and receiver bandwidth and that the sensitivity increased with spatial resolution. Materka and Strzelecki [56], in a study of the effects of magnetic field bias on texture features, found that texture features could be sensitive to inhomogeneity because the variation in intensities could obscure the underlying texture. The same group also reported the robustness of some gray-level co-occurrence matrix features (compared with other texture metrics) to magnetic field–based nonuniformities.

Multiple studies have assessed the reliability of PET-based radiomics. Leijenaar et al. [57], in a radiomics study of non–small cell lung carcinoma, reported that most of the features assessed had both high test-retest reliability (71%) and high interobserver agreement (91%). In general, the radiomic features that were more stable on repeated PET examination were also found to be more robust against interobserver variability in segmenting the lesion. In a study of esophageal cancer, Tixier et al. [58] reported that some Haralick texture metrics, such as entropy, dissimilarity, and homogeneity, had comparable or better reproducibility (mean difference, < 6%) when compared with simple PET measurements, such as the standardized uptake value. Various studies showed large variations (> 30%) when the number of iterations, grid size, reconstruction algorithm, postreconstruction filter, or a combination of these elements were changed in their study of PET examinations of solid tumors [59–61].

No studies assessing the reliability of ultrasound-based radiomics studies were found. However, an isolated study by Gómez et al. [11], which analyzed co-occurrence texture metrics as a function of gray-level quantization for classifying breast ultrasound findings, showed that, without averaging, the quantization levels do not impact the discrimination power (AUC value, 87). However, with averaging, the quantization levels negatively impact the discrimination power (AUC value, 81).

The Path Forward

For radiomics to be translated into practice and eventually be accepted in selected areas in clinical practice, numerous challenges need to be addressed [43]. These challenges are presented in the following subsections.

Variability in Imaging and Imaging Processing Protocols

A wide variety of imaging scanners are available in different centers. It is not practical to have uniform protocols across different hospitals and imaging centers. As an alternative to the standardization of preacquisition protocols, postacquisition harmonization corrections can be applied to data collected across the different centers [62, 63]. Mackin et al. [63] showed that by applying a correction based on resampling and using a Butterworth low-pass filter in the frequency domain, variability in CT radiomic features caused by variations in pixel size can be reduced. Orlandi et al. [62] used gaussian smoothing protocols to harmonize data from three different PET scanners to reduce multicenter variability for measuring textural features and standardized uptake values. Phantom studies have been used to assess the reliability of radiomic metrics, with variations in the radiomic process limited to texture [45, 48, 64, 65]. A more comprehensive assessment of all types of radiomic features, including size and shape, is needed to assess comparative performance. Future studies should provide information regarding imaging parameters and image processing steps, to reliably represent their results. Details regarding the use of denoising, artifact removal, and data transformation should be justified and described as outlined in the Materials and Methods section.

Problems Caused by Data Misrepresentation

An additional factor that impedes thorough evaluation of classifiers is data misrepresentation resulting from the small sample sizes associated with biomedical data [66]. Researchers must try to ensure class balance (i.e., a comparable sample size in all phenotypic groups) in both the training and test dataset. Currently, data sampling methods such as random upsampling, downsampling, and other methods are being used to balance the data [67]. Depending on the characteristics of the imbalanced dataset, the optimal solutions will vary. Ideally, to avoid information leakage, the training and test dataset should be predetermined before the start of the study. Although techniques such as the leave-one-out method are being used to evaluate classifier performance in smaller datasets, the generalizability of the result is drastically limited compared with that of studies in which independent cohorts are used for testing [68]. Future studies should emphasize reporting performance in independent cohorts.

Variability in ROI Segmentation

Manual segmentation introduces inter- and intraobserver variability. Automated techniques, however, have to be developed for individual organ systems, and accuracy is not yet established. At present, manual or semiautomatic segmentations are used for radiomics, with reasonable to good intra- and interoperator variability noted (< 15% variability) [52, 69]. However, this finding is dependent on results from particular types of studies, such as CT of thoracic lesions, and cannot be generalized to all studies. Studies adopting manual segmentation must report intra- and interobserver agreement results to show the generalizability of the approach. In addition, the selection of a tumor ROI from which the texture metrics are derived is critical to the assessment of intratumoral heterogeneity. It also is not yet established whether the entire tumor volume or a representative tumor section should be taken for the selection of ROI [70, 71]. Future studies in which 3D volumetric data are acquired should compare the performance of the 2D ROI-based texture metrics to volumetric ROI-based texture metrics.

Variability in Feature Extraction

Features have to be generated that reliably reflect the complexity of the individual volumes but cannot be overly complex or redundant. Currently, numerous techniques and algorithms are used, although no consensus exists regarding a standard method [72, 73]. Freeware, such as IBEX [74], PyRadiomics [33], ImageJ [75], and MaZDa [76], is available to the community to advance radiomics research; however, studies show a discrepancy in results on the basis of the algorithms used [77]. In a repeatability study, Foy et al. [77] reported excellent agreement among all first-order texture metrics, except for kurtosis, with use of different software programs. In the same study, second-order metrics showed moderate to poor agreement. Such studies show that discrepancies exist in each program that could potentially result in variability in the resultant texture features. Future radiomics studies should account for and detail methods to overcome such differences when using open-source software.

Variability in Feature Selection Methods

The next step after feature extraction involves feature selection or reduction to optimize classifier performance on the given training datasets. A training set consists
of samples to train the candidate classifiers; cross-validation is often used to select the best classifier among candidate classifiers. Feature reduction is an important step to avoid overfitting the classifier to training data, especially when the number of training samples is significantly less than the number of extracted features. Multiple statistical measures exist to reduce the number of features to optimize the performance of candidate classifiers, such as decision tree, neural network, Bayesian, random forest, multivariate adaptive regression, and other classifiers [78]. However, each statistical measure may provide a different list of features, depending on the inherent search criteria and evaluation method used [79]. As a result, the performance of resultant classifiers will vary depending on the final list of features.

**Variation in Classifier Validation Metrics**

Once trained, the performance of candidate classifiers is validated using independent testing data. Depending on the choice of cross-validation technique and assessment metric used, the final classifier selected for a given application will vary. The AUC value is one of the most common methods used to assess the accuracy of predictive distribution models. However, the AUC value can be limited as a performance predictor [80], particularly in scenarios in which the data are skewed, thereby undermining the capability to realistically assess classifier performance. Therefore, researchers should use multiple metrics (AUC, false-positive, true-positive, recall, precision, and F-measure values) to select the final pool of classifiers [66]. The performance of the final pool of classifiers is then evaluated using test data (i.e., independent data not used for training or validation); typically, the test data distribution should not vary significantly from the training data distribution. The classifier that performs best with the test data is selected as the final classifier for the given clinical application.

As the field of radiomics advances, it will be critical to have a consortium of researchers addressing radiomics using diverse imaging modalities and research questions. The goal of the consortium would be to create consensus-based guiding principles or guidelines for the radiomics workflow. In addition, providing open access to data and programming code will help promote improved reliability and further provide opportunities for conducting large-scale radiomics studies us-

### TABLE 2: Checklist of Suggested Attributes in Future Radiomics Studies

<table>
<thead>
<tr>
<th>Module</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of study</strong></td>
<td>Provide general details of the study</td>
</tr>
<tr>
<td>Retrospective versus prospective, sample size, power calculation, data source (single institution or multiple institutions), and single or multiple scanners; similar or differing scan protocols</td>
<td></td>
</tr>
<tr>
<td><strong>Image acquisition</strong></td>
<td>Provide technical transparency</td>
</tr>
<tr>
<td>Imaging modality (with or without contrast enhancement), multiphase or dynamic studies, different scan manufacturers, and image acquisition and processing protocol details (including noise and artifact removal)</td>
<td></td>
</tr>
<tr>
<td><strong>Image segmentation</strong></td>
<td>Provide technical transparency of the approach</td>
</tr>
<tr>
<td>Included ROIs, excluded regions, manual segmentation or automated or semiautomated segmentation methods, 2D or 3D segmentation, image registration algorithm, voxel interpolation methods, and gray-level discretization methods</td>
<td></td>
</tr>
<tr>
<td><strong>Feature extraction</strong></td>
<td>Provide technical transparency of the algorithms used</td>
</tr>
<tr>
<td>Details regarding equations; documented code (open-source or in-house code); information regarding code and data access, if applicable; details regarding preextraction filters; postprocessing steps, and other tasks</td>
<td></td>
</tr>
<tr>
<td><strong>Statistical analysis</strong></td>
<td>Provide details of the statistical design depending on the approach</td>
</tr>
<tr>
<td>I. Use of radiomic feature as the metric in predicting disease type or treatment outcome</td>
<td></td>
</tr>
<tr>
<td>A. Pilot study with the purpose of providing details regarding the following categories</td>
<td></td>
</tr>
<tr>
<td>1. Performing measurement reliability check</td>
<td></td>
</tr>
<tr>
<td>2. Controlling for multiple testing error</td>
<td></td>
</tr>
<tr>
<td>B. Confirmative study with the purpose of providing details regarding the following categories</td>
<td></td>
</tr>
<tr>
<td>1. Performing measurement reliability check</td>
<td></td>
</tr>
<tr>
<td>2. Reserving independent testing data</td>
<td></td>
</tr>
<tr>
<td>3. Performing variable selection &amp; cross-validation using training data</td>
<td></td>
</tr>
<tr>
<td>II. Use of radiomic feature as the outcome</td>
<td></td>
</tr>
<tr>
<td>A. Pilot study with the purpose of providing details regarding the following categories</td>
<td></td>
</tr>
<tr>
<td>1. Performing measurement reliability check</td>
<td></td>
</tr>
<tr>
<td>2. Performing data normality check</td>
<td></td>
</tr>
<tr>
<td>3. Controlling for multiple testing error</td>
<td></td>
</tr>
<tr>
<td>B. Confirmative study with the purpose of providing details regarding the following categories</td>
<td></td>
</tr>
<tr>
<td>1. Performing measurement reliability check</td>
<td></td>
</tr>
<tr>
<td>2. Performing data normality check</td>
<td></td>
</tr>
<tr>
<td>3. Performing variable reduction</td>
<td></td>
</tr>
<tr>
<td>4. Mixed effect model for longitudinal data (if applicable)</td>
<td></td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>Provide details regarding the validation approach</td>
</tr>
<tr>
<td>Details regarding the sample size of the independent data, the source of data, single institution or multiple institutions, single scanner or multiple scanners, and which stages of the radiomic analysis were blinded. If analysis was not blinded, identify potential stages of bias.</td>
<td></td>
</tr>
</tbody>
</table>
Varghese et al.


Texture Analysis of Imaging


72. Dinapoli N, van Soest J, Maschiochi C, et al. Radiomics in magnetic resonance imaging for prognosis in patients with rectal cancer: an inde-


75. Abramoff MD, Magalhães PJ, Ram SJ. Image processing with ImageJ. *Biophotonics Int* 2004; 11:36–42


93. Varghese BA, Chen F, Hwang DH, Cen SY, Gill IS, Duddalwar VA. Differentiating solid, non-macroscopic fat containing, enhancing renal masses using fast Fourier transform analysis of multiphase CT. *Br J Radiol* 2018; 91:20170789

94. Al-Kadi OS. Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed tomography lung tumour images. *Comput Med Imaging Graph* 2010; 34:494–503


This article has been cited by:
