TOMOGRAPHY® RESEARCH ARTICLE

Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Data Sets

M. McNitt-Gray[1], S. Napel[2], A. Jaggi[2], S.A. Mattonen[2,3], L. Hadjiiski[4], M. Muzi[5], D. Goldgof[6], Y. Balagurunathan[7], L.A. Pierce[5], P.E. Kinahan[5], E.F. Jones[8], A. Nguyen[8], A. Virkud[4], H.P. Chan[4], N. Emaminejad[1], M. Wahi-Anwar[1], M. Daly[1], M. Abdalah[7], H. Yang[9], L. Lu[9], W. Lv[10], A. Rahmim[10], A. Gastounioti[11], S. Pati[11], S. Bakas[11], D. Kontos[11], B. Zhao[9], J. Kalpathy-Cramer[12], and K. Farahani[13]

[1]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA; [2]Stanford University School of Medicine, Stanford, CA; [3]The University of Western Ontario, Canada; [4]University of Michigan, Ann Arbor, MI; [5]University of Washington, Seattle, WA; [6]University of South Florida, Tampa, FL; [7]H. Lee Moffitt Cancer Center, Tampa, FL; [8]UC San Francisco, School of Medicine, San Francisco, CA; [9]Columbia University Medical Center, New York, NY; [10]BC Cancer Research Centre, Vancouver, BC, Canada; [11]Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA; [12]Massachusetts General Hospital, Boston, MA; and [13]National Cancer Institute, Bethesda, MD

Corresponding Author:
Michael McNitt-Gray, PhD
Department of Radiological Sciences, David Geffen School of Medicine at UCLA, 924 Westwood Blvd, Suite 650, Los Angeles, CA 90024;
E-mail: mmcnittgray@mednet.ucla.edu

Key Words: Radiomics, Quantitative Imaging, Standardization, Multi-center, Feature Definitions

Abbreviations: Computed tomography (CT), poitron emission tomography (PET), digital imaging and communications in medicine (DICOM), region of interest (ROI), volume of interest (VOI), quantitative imaging network (QIN), digital reference object (DRO), three dimensional (3D), two dimensional (2D), lung image database consortium (LIDC), coefficient of variation (CV), response evaluation criteria in solid tumors (RECIST), lung CT screening reporting & data system (Lung-RADS), Hounsfield Units (HU), gray level co-occurrence matrix (GLCM), field of view (FOV), quantitative image feature engine (QIFE), quantitative image analysis (QIA), standardized environment for radiomics analysts (SERA), cancer imaging phenomics toolkit (CaPTK)

ABSTRACT

Radiomic features are being increasingly studied for clinical applications. We aimed to assess the agreement among radiomic features when computed by several groups by using different software packages under very tightly controlled conditions, which included standardized feature definitions and common image data sets. Ten sites (9 from the NCI's Quantitative Imaging Network] positron emission tomography–computed tomography working group plus one site from outside that group) participated in this project. Nine common quantitative imaging features were selected for comparison including features that describe morphology, intensity, shape, and texture. The common image data sets were: three 3D digital reference objects (DROs) and 10 patient image scans from the Lung Image Database Consortium data set using a specific lesion in each scan. Each object (DRO or lesion) was accompanied by an already-defined volume of interest, from which the features were calculated. Feature values for each object (DRO or lesion) were reported. The coefficient of variation (CV), expressed as a percentage, was calculated across software packages for each feature on each object. Thirteen sets of results were obtained for the DROs and patient data sets. Five of the 9 features showed excellent agreement with CV $< 1\%$; 1 feature had moderate agreement (CV $< 10\%$), and 3 features had larger variations (CV $\geq 10\%$) even after attempts at harmonization of feature calculations. This work highlights the value of feature definition standardization as well as the need to further clarify definitions for some features.

INTRODUCTION

Radiomics is described as the high-throughput extraction of large amounts of image features from radiographic images (1–4). Radiomic features provide quantitative descriptions of objects (tissues, suspected pathology, and anatomic regions) contained within the image data. These mathematical descriptors provide ways to characterize the size, shape, texture, intensity, margin, and other aspects of the imaging features of these objects, with

ISSN 2379-1381 https://doi.org/10.18383/j.tom.2019.00031

the eventual goal of being able to accomplish a diagnostic imaging task, such as distinguishing benign from malignant nodules, assessing response to therapy, correlating imaging with genomics, or decoding the object's imaging phenotype and predicting survival outcomes (5). Within the NCI's Quantitative Imaging Network (6), there are concerted efforts to bring quantitative imaging and radiomic features into clinical trials to aid in treatment evaluation. Among specific efforts have been investigations into these clinical tasks as well as the sensitivity of extracted radiomic feature values to various aspects of the quantitative imaging chain such as image segmentation (7–8).

Despite the widespread use of radiomic features and even the public availability of software packages to compute radiomic features (9–12), there has been a lack of consistency in the definition of radiomic feature values and how they are calculated. This may be a factor contributing to the lack of reproducibility of results across different data sets and different institutions and specifically across different software packages that perform radiomic feature calculations. Recently, there has been an effort to standardize feature definitions as described by the Image Biomarker Standardization Initiative (IBSI), which has published a reference manual (13). This reference manual has described a large number of features in detail and has even introduced conventions that provide unique codes to identify each of these features.

Therefore, the primary purpose of this study was to investigate the level of agreement in radiomic feature values that could be achieved across a set of institutions and software packages when using a set of common feature definitions, common image data sets (digital reference objects and patient data sets) and common object definitions (segmentations). Ideally, under these controlled conditions, all institutions and software packages would obtain the same values for all radiomic features for all objects. A secondary purpose of this study was to identify the underlying issues when that did not occur (when there were feature value disagreements) and to address those issues when possible. It should be pointed out that the goal of this study was to neither determine the utility of the identified features nor identify superiority of a single tool, in any specific clinical task. That is, this study does not evaluate which features demonstrate efficacy in predicting whether a lung lesion is benign or malignant or whether a patient is responding to therapy. This study also does not assess how sensitive various features are to different sources of variability in the quantitative imaging chain such as object segmentation or image acquisition and reconstruction parameters; these issues are being addressed in other efforts (7, 14–17). The primary goal is to ascertain how much detail is necessary to be reported on future studies to sufficiently describe a feature such that investigators at different institutions and/or using different software packages can produce consistent results. We consider this to be a very important step in the process of developing robust radiomic features that will ultimately contribute to the use of quantitative imaging methods in clinical trials and clinical practice.

METHODOLOGY

Overview

This project was initiated by the positron emission tomography–computed tomography working group of the National Cancer Institute's Quantitative Imaging Network. Ten sites participated in the investigation, which took place in 4 phases. In the first phase, a small subset of radiomic features were identified that would make this project feasible to pursue with some depth. In the second phase, radiomic feature values were calculated by participating sites on a set of digital reference objects (DROs) to help identify potential issues. In the third phase, radiomic feature values were calculated by participating sites for identified objects in a small set of patient image data sets [identified lung nodules from the Lung Image Database Consortium (LIDC) data set (18)]. In the fourth phase, an effort was made to specifically harmonize the methodology for calculating one of the more complex radiomic features and to determine if the level of agreement could be improved. Each of these phases is described in more detail in the following sections.

Features Selected and Their Definitions

Investigators involved in the project agreed to identify 9 radiomic features to be investigated for this project. This subset of features was selected to simultaneously keep the number of features to a manageable level for this project while including features from several key categories including morphology (size and shape), intensity, and image texture. The IBSI reference manual (13) was identified as a reference, and feature definitions would be consistent with the definitions and conventions of that resource. Note that where the IBSI reference manual is cited, the section numbers and codes provided are from Version 10.

All features are defined for an identified object whose boundary is the volume of interest (VOI) accompanying each DRO or patient image data set. The 9 specific features selected for analysis were:

1. **Approximate volume** (IBSI Section 3.1.2; code YEKZ): This feature is a commonly used size descriptor that counts the number of voxels within an identified object and multiplies by the voxel size in cubic millimeter. Therefore, the volume is expressed in cubic millimeter.
2. **2D diameter**: Although this feature is commonly used in oncology assessments such as RECIST and Lung-RADS (19), it is not described in the IBSI reference manual. For this investigation, we agreed to calculate the diameter in a single axial slice. However, sites were allowed flexibility in both how the slice was selected and in how the longest diameter was calculated. For example, some sites calculated the longest chord from all boundary points on each slice and chose the largest of these as the 2D diameter. Others calculated the longest chord on the slice with the largest area and chose this as the 2D diameter. In all cases, the 2D diameter was specified in millimeter.
3. **3D diameter** (IBSI Section 3.1.11, code L0JK): This feature is the distance between the 2 most distant vertices from the set of boundary voxels in the VOI. These vertices are not constrained to lie in the same image plane. In this project, the diameter was specified in millimeter.
4. **Mean intensity** (IBSI Section 3.3.1, code Q4LE): This feature is the mean intensity value over the VOI. In this project, because the patient image data sets were computed tomography (CT) scans, and the DROs were scaled apropos to CT scans, the mean intensity was specified in Hounsfield Units (HU).

5. **Standard deviation**: This feature is not explicitly defined by the IBSI reference manual; however, it does define the "intensity variance" (Section 3.3.2, code ECT3). In this project, the standard deviation in HU was defined as the square root of the intensity variance.
6. **Kurtosis** (IBSI Section 3.3.4, code IPH6). This feature is technically referred to as the excess kurtosis and is based on the fourth moment of the intensity distribution and is used as a measure of peakedness in that intensity distribution. It should be noted that in this definition, kurtosis is corrected by a Fisher correction of −3 to center it on 0 for normal distributions. This feature is dimensionless.
7. **Surface area** (IBSI Section 3.1.3, code C0JK): This feature is calculated based on the mesh approach described in IBSI Section 3 and specifically from a VOI mesh defining the surface of the object by summing over the triangular face surface areas (mm^2) of the mesh.
8. **Sphericity** (IBSI Section 3.1.8, code QCFX): This dimensionless feature describes how sphere-like the volume is. It is based on the ratio between volume (calculated from the mesh used for surface area, not the approximate volume feature used in this study) and surface area.
9. **Gray Level Co-occurrence Matrix entropy** (IBSI Section 3.6.4, code TU9B): This feature is described as "Joint Entropy" by the IBSI reference manual and is a texture measure described by Haralick et al. (20) using the Gray Level Co-occurrence Matrix (GLCM) approach. In the IBSI reference manual, there are descriptions of how the matrices are formed and how the features are aggregated (6 different methods for aggregation are described). In the first phases of this project, sites used whatever their local software package allowed in terms of aggregation and other parameters. In the final phase of this project, there was an effort to harmonize the approaches and parameters for this feature. This will be described in more detail in the following sections. This feature is dimensionless.

Image Data Sets

This study used 2 different image data sets. The first data set was a set of DROs created by the participants from Stanford (21). The second data set was a set of 10 patient image studies from the publicly available Lung Image Database Consortium (18) collection hosted on the The Cancer Imaging Archive (TCIA) website (22). For each data set (DRO and patient data sets), we identified a specific object for use in this project. One VOI was created for each object and all sites used that same VOI definition. Based on our experience with a previous project (7), both DICOM and NIfTI formats were created for each image data set and the VOI was provided in each format as well (DICOM Segmentation Object (DSO) as well as NIfTI segmentation boundary). All image data (DRO and patient image data) as well as VOIs in both DSO and NIfTI formats are publicly available at https://doi.org/10.7937/tcia.2020.9era-gg29.

DROs. The DROs were created by the participants at Stanford and served as a starting point for our analysis (21). These DROs are generated from continuous 3D functions with known "features" (eg, volumes, mean intensities) that are sampled and stored as DICOM (or NIfTI) images. These features have settable parameters and known definitions as specified in (21). For this project, the DROs were created with a 512×512 matrix similar to typical CT images and used a 500-mm display FOV. The slice thickness and spacing were each set to 1.0 mm. This results in voxels that are $\sim 1.0 \times 1.0 \times 1.0$ mm. The intensity of the DROs were defined using mathematical functions that set the internal gray values. These values can be converted to specific intensity values (eg, HU for CT) when desired. The DROs used in this study used the DICOM fields of rescale intercept [(0028, 1052) with a value of −1024] and rescale slope [(0028, 1053) with a value of 1] similar to CT images. Saved as conventional medical images, these objects can have the same radiomics operations applied to them as to patient images, allowing the accuracy of feature calculations to be determined in a controlled fashion. This is of substantial value when trying to achieve the same feature values across software packages, and these objects play a vital role in helping sites diagnose underlying issues.

Figure 1 shows the 3 DROs used in this study; each was designed to exercise different radiomic features. The DROs created for this project were: (1) a mathematically defined sphere with a radius of 100 mm and with uniform intensities (100 HU) for all voxels within the VOI; (2) a mathematically defined sphere with a radius of 100 mm, but having *intensity variation* that is sinusoidal in 3 dimensions with a mean of 100 HU, an amplitude of 50 HU, and a wavelength of 10 mm (3) a uniform intensity (100 HU) object with a *nonspherical shape* that is described by a *sinusoidal variation in the location of the surface* with a mean radius of 100 mm, an amplitude of 20 mm, and unitless azimuthal and inclination angular frequencies of 9 (ie 9 "bumps" in a cross section). More detailed descriptions are available in the study by Jaggi et al. (21), and the DROs are publicly available in DICOM and NIfTI formats at: https://doi.org/10.7937/tcia.2020.9era-gg29 .

Patient Data Sets. The patient data sets for this study were a subset of patient data sets used in a previous study (7, 14). Specifically, the same 10 cases selected from the LIDC data set (18) that were used in that previous study were used in this study (see details of cases in the online supplemental Appendix). As in that previous study, a single lesion from each case was identified for analysis. That previous study generated VOIs using algorithms from 3 academic institutions and each method performed three repeat runs on each nodule. For this study, and to eliminate one source of variability, 1 VOI was created for each lesion and all sites used that same VOI definition. The specific VOI chosen for each lesion was the first run of the first algorithm [Algorithm 1 in Kalpathy-Cramer et al. (14)]. Based on our experience with that previous project (7), both DICOM and NIfTI formats were created for each image data set, and the VOI was provided in each format as well (DSO and NIfTI segmentation boundary).

Three example lung lesions are shown in Figure 2. The online supplemental Appendix contains a table with details about the lesions including a description of which LIDC data set was used and the range of nodule sizes and densities (1 nodule was clearly calcified). That table also shows that 5 of the 10 cases used contiguous reconstructions (slice spacing = slice thickness), while the other 5 used overlapped reconstructions (slice spacing < slice thickness). Finally, this table also shows that images acquired from different scanners (GE vs Philips) used different values for

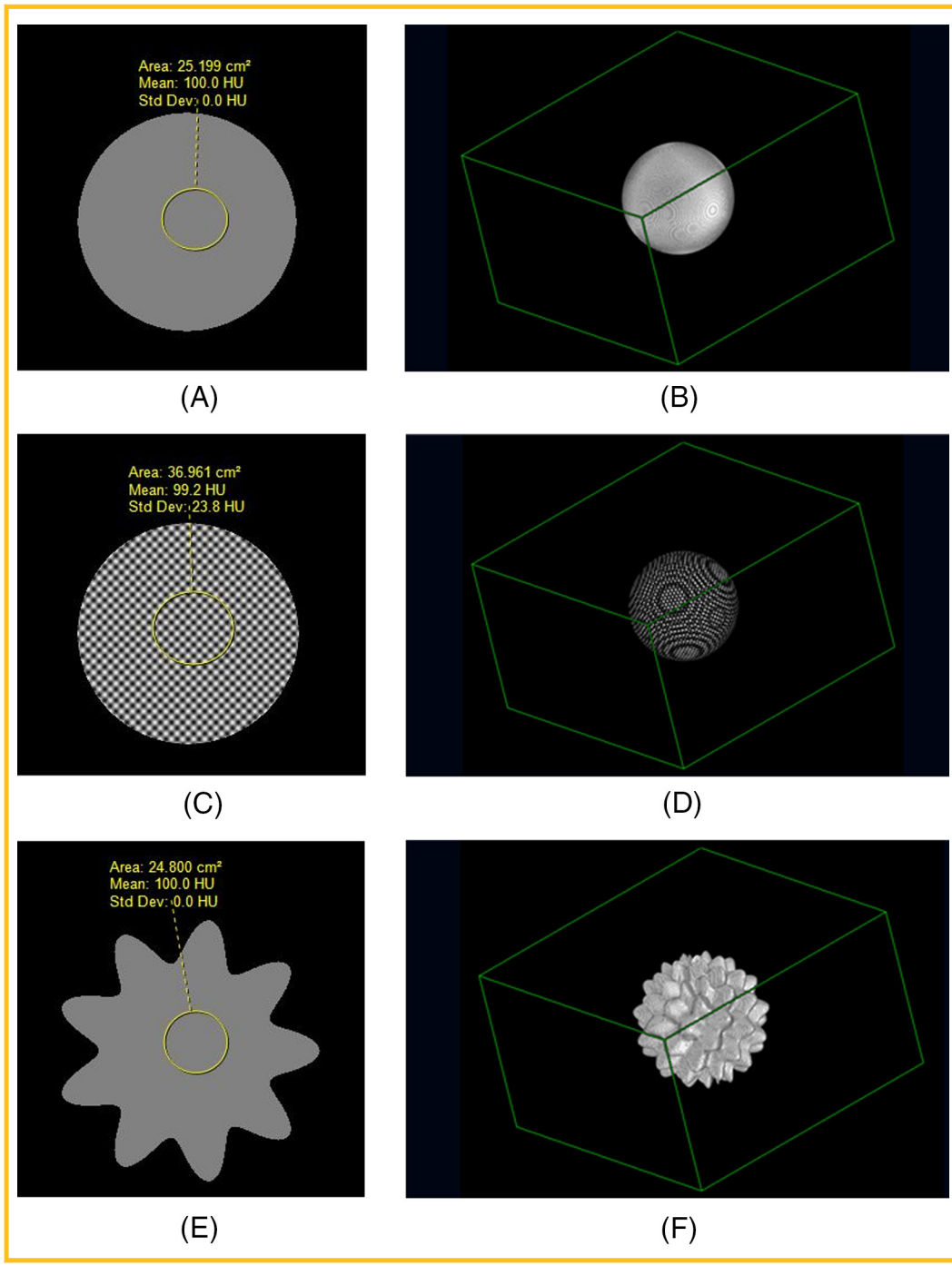

Figure 1. Digital reference objects (DROs) used in this study: axial cross section of sphere with uniform intensity (A) and 3D rendering of the sphere (B). Axial cross section of sphere with intensity variation (C) and 3D rendering of the object (D); nonspherical (but mathematically defined) object with uniform intensity (E) and 3D rendering of this object (F).

the conversion from gray levels to HU (DICOM tag rescale intercept; image data sets from GE scanners used −1024 as the rescale intercept, while image data from Philips scanners used rescale intercept of −1000).

Software Packages Used

Feature calculation results were received from 10 different sites. Some sites submitted results from 2 different software packages. The software packages were primarily developed by the participating institutions with the exception of PyRadiomics (10, 23) which is an open source Python-based radiomics package that 3 different sites used (1 used only PyRadiomics and 2 other sites submitted results from their own software as well as data from PyRadiomics). Table 1 summarizes the software packages used in this study as well as whether they used DICOM, DSO, or NIfTI formats for the images and VOIs.

Harmonization of GLCM Entropy

Each site and software package (including the authors of PyRadiomics) was surveyed to determine if they were calculating radiomic features according to the definitions described by the IBSI (above). In some cases, the software packages were developed by others and it could not be precisely determined how the features were being calculated. In

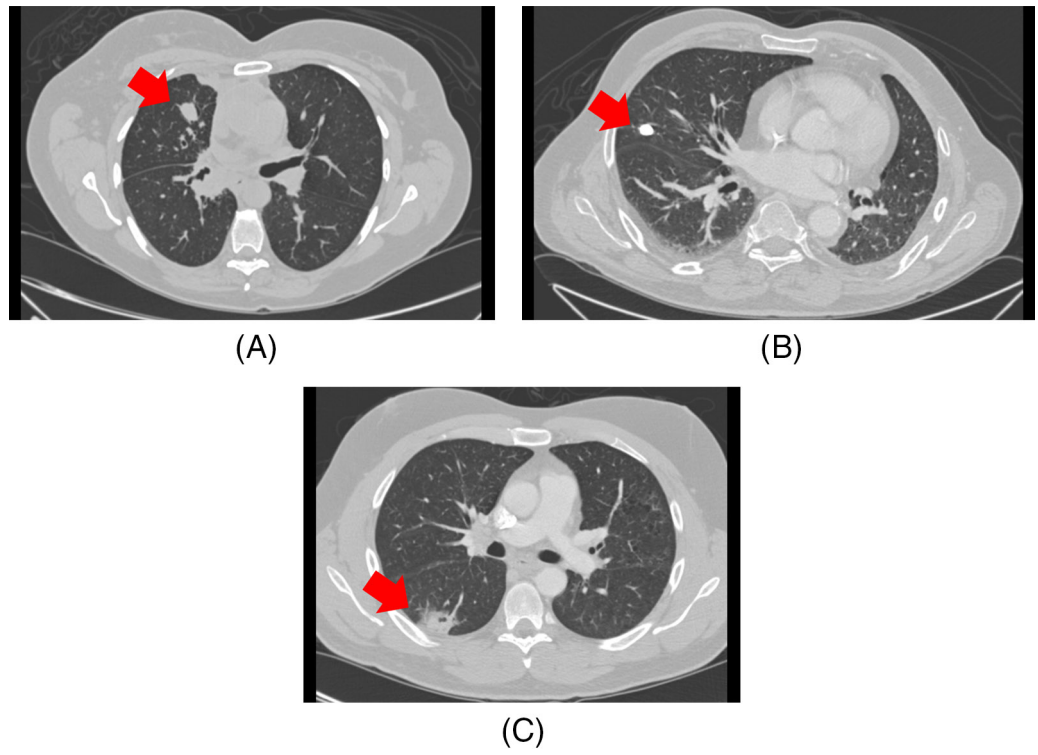

Figure 2. Examples showing cross sections of 3 of the 10 cases used in this investigation: from LIDC-0314 with a solid lesion having a diameter of ~17 mm (A); from LIDC-0766 with a calcified lesion having a diameter of ~13 mm (B); and from LIDC-0811 with a part-solid lesion having a diameter of ~30 mm (C).

the specific case of the GLCM entropy feature, there are several choices in how this feature was calculated and this was determined as best as possible. This included trying to determine parameters related to the formation of the co-occurrence matrix (distance, angle), quantization of gray levels (HU), the number of levels to form the matrix, and whether the implementation required both the source and destination voxel, or just the source voxel, to be contained within the VOI. Other parameters surveyed were related to the calculation of the feature including how the feature aggregation was being performed [in Section 3.6 of the IBSI reference manual (13)].

The survey of the participating sites and their software packages revealed that there was a wide range of approaches to calculating GLCM entropy. The initial decision was made to proceed with whatever default parameters were being used in each site's software package(s) for the purposes of assessing agreement using the DROs and patient data sets. The decision was also subsequently made to determine if sites could harmonize their calculation of the GLCM entropy feature values by using as many

Table 1. Information About Software Packages Used in This Study

Site	Software	Image Data Used	VOI Data Used
1. Stanford	Quantitative Image Feature Engine (QIFE) (9, 24)	DICOM	DSO
1. Stanford	PyRadiomics (10, 23)	DICOM	DSO
2. UCLA	Quantitative Image Analysis (QIA) (25–27)	DICOM	NIfTI
2. UCLA	PyRadiomics (10, 23)	DICOM	NIfTI
3. UW	PMOD (28)	DICOM	NIfTI
3. UW	PORTS (GLCM only) (29)	DICOM	NIfTI
4. USF	Package 1	DICOM	NIfTI
5. Moffit	Package 2	DICOM	NIfTI
6. Columbia	Package 3 (17)	DICOM	NIfTI
7. Michigan	MiViewer (30, 31)	DICOM	NIfTI
8. BC Cancer	SERA (32)	DICOM	NIfTI
9. Penn (CBICA)	CaPTK (11)	DICOM	NIfTI
10. UCSF	PyRadiomics (10, 23)	DICOM	DSO

Table 2. Coefficient of Variation Results[a]

DRO	Approximate Volume	Surface Area	2D Diameter	3D Diameter	Sphericity	Mean Intensity	Standard Deviation	Kurtosis[b]	GLCM Entropy
Uniform Phantom	0.004%	13.41%	0.23%	0.27%	12.82%	0.00%	—	—	1002%
Intensity Varying Phantom	0.004%	13.41%	0.23%	0.27%	12.82%	0.00%	0.11%	0.31%	50.9%
Shape Varying Phantom	0.010%	12.27%	0.71%	0.18%	11.70%	0.00%	—	—	625%

[a] CV results are expressed as a percentage for each feature and each DRO across all 13 submissions. Note that there was no value for standard deviation and kurtosis for the uniform phantom and shape varying phantom, as the intensity values for these phantoms were all set to the same value (100 HU).
[b] w/Fisher correction.

common parameter settings (without rewriting any code) as possible.

Specifically, sites harmonized their calculation of GLCM entropy by using the following in their calculations:

1. Interpolate to isotropic 1-mm voxels.
2. Calculate the features in 3D with 13 angles and d = 1 mm.
3. Select 256 gray levels for quantization and use the voxels just within the VOI for this quantization step (rather than the entire image volume).
4. For the aggregation method, it seems that most software packages were fixed, but for PyRadiomics, there were a few options that could be selected. One was that features were computed from each 3D directional matrix and then averaged over the 3D directions to arrive at a single feature value. Another option was that the feature value is computed from a single matrix after merging all of the 3D directional matrices. These correspond to options (e) "volume without merging" and (f) "volume with full merging," as described in Figure 3.3 of the IBSI reference manual (13).

It should be noted here that these parameter settings were not determined to be optimal in any way. Instead they were identified as the parameter settings most likely to be common among the software implementations used by our investigators and therefore most likely to lead to values that would be agreed upon by different sites and software packages.

Reporting Results

Results of feature calculations were reported for 3 distinct phases of this project: (a) DROs, (b) patient data sets, and (c) harmonized GLCM entropy results for all patient data sets.

For the DROs, each site reported the value of each of the 9 indicated features for each DRO phantom using the indicated software package. The result was that for each feature calculated on each DRO, the mean value, standard deviation, and coefficient of variation (CV) (expressed as a percentage) were calculated across software packages. For the patient data sets, a similar approach was used so that for each feature calculated on each patient data set, the mean value, standard deviation, and CV (expressed as a percentage) were calculated across software packages. For the harmonized GLCM entropy phase of the project, only the GLCM entropy value using the harmonized set of parameters was recorded for each patient image data set for each software package; the mean value, standard deviation and CV (expressed as a percentage) were calculated across software packages.

RESULTS

The results from the 3 investigations (feature calculations performed on DROs, feature calculations performed on patient studies, and GLCM feature calculations after using the harmonized parameter settings described in the previous section) are described in the following sections.

DRO Results

Results were received from 13 different submissions from 10 sites for this part of the study; as indicated in Table 1 above, 2 sites submitted results from their own locally developed software as well as results from PyRadiomics, and 1 group submitted results from their own locally developed software as well as a software package called PMOD. Table 2 shows the results of the CV expressed as percentage (CV%) by phantom and by feature.

Six of the 9 features calculated show excellent agreement across submissions and phantoms with CV $<$ 1%; specifically, the approximate volume, 2D diameter, 3D diameter, mean intensity, standard deviation, and kurtosis (after Fisher adjustment) features. Larger variations (12%–13%) were observed for the surface area and sphericity features. Finally, extremely large variation values were observed (51%–1001%) for the GLCM entropy feature across submissions and phantoms. One note about the GLCM entropy variation is that in the uniform phantom and the shape varying phantom, the intensity values within the VOI were completely uniform and therefore most software packages returned a GLCM entropy value of 0 (or very close to 0). However, 2 submissions (out of 13) had nonzero GLCM entropy values for these 2 phantoms, resulting in a very small mean value across submissions (0.1 for the uniform phantom and 0.2 for the shape varying phantom) which led to a very large CV.

Patient Data Set Results

For the patient data sets, 13 different submissions originating from 10 sites were also received. Table 3 shows the mean and

Table 3. Mean and Standard Deviation of the Coefficient of Variation Values[a]

Coefficient of Variation	Approximate Volume	Surface Area	2D Diameter	3-D Diameter	Sphericity	Mean Intensity	Standard Deviation	Kurtosis[b]	GLCM Entropy
Mean CV	0.00%	17.06%	8.44%	3.27%	16.90%	0.00%	0.07%	−0.44%	36.24%
Standard Deviation CV	0.00%	2.49%	5.19%	2.25%	3.57%	0.00%	0.09%	0.74%	4.66%

[a] The values are expressed as a percentage, for each feature calculated across all 10 patient nodule cases and all 13 submissions.
[b] w/Fisher correction.

standard deviation of the CV as percentage (CV%) values for each feature. This was calculated by first calculating the CV for each patient data set across software packages and then calculating the mean and standard deviation of those CVs across all 10 patient cases.

These results show that only 4 of the 9 features calculated show the same level of excellent agreement (CV $<$ 1%) as observed with the DROs; specifically, the approximate volume, mean intensity, standard deviation, and kurtosis (after Fisher adjustment) features. Slightly larger variations (mean CV of 3%–9%) were observed for 2D diameter and 3D diameter features. Larger variations (mean CV = 17%) were again observed for the surface area and sphericity features. Finally, larger variations were observed (mean CV = 37%) for the GLCM entropy feature across submissions and patient cases. Note that because these were patient data sets and the intensity values were not completely uniform, the mean GLCM entropy values for each patient data set actually ranged from 5.4 to 7.8 (with a much wider range of individual values from different software packages); this in turn led to smaller CV values than those observed for uniform DROs.

Table 4. Coefficient of Variation Results for the GLCM Entropy Feature Values

Case	CV (%) Harmonized Settings	CV (%) Default Settings (Table 3 results)
1	18.25%	41.80%
2	18.02%	38.55%
3	20.05%	30.89%
4	30.34%	43.57%
5	15.15%	33.93%
6	21.69%	37.08%
7	14.49%	39.37%
8	25.69%	29.98%
9	14.06%	35.70%
10	17.51%	31.55%
Mean	19.52%	36.24%
Standard Deviation	5.19%	4.66%

The values are expressed as a percentage for each case (across submissions) as well as the mean and standard deviation across cases. The first column reports the results when using the harmonized parameters described above. The second column reports the results when using the default (non-harmonized) parameters.

Results from GLCM Entropy Harmonization

For this final investigation, 11 submissions from 9 sites were received in which the harmonized parameter settings for the GLCM entropy calculation (described above) were used. Table 4 shows the individual CV values for each case (calculated across submissions) under both the harmonized parameter settings and each site's default (nonharmonized) settings used to obtain the results in the previous section. Table 4 also shows the mean and standard deviation of the CV as percentage (CV%) values for each feature across cases (and submissions).

These results show a substantially reduced CV when the parameters are harmonized across software packages compared with CV when default settings are used (mean CV of 19.3% vs 36.2%). This is true for each individual case as well. However, it should also be noted that the agreement is still relatively modest (mean CV% ~20% across cases and software packages), which indicates there are still outstanding issues to resolve to obtain the levels of agreement seen in the features with the excellent levels of agreement (eg, approximate volume, mean intensity, etc.).

DISCUSSION

The purpose of this work was to investigate the level of agreement among radiomic features that could be achieved when computed by several groups using different software packages. A secondary purpose of this work was to use these investigations to identify issues that led to differences in feature values produced by software packages and to determine if these issues could be readily addressed.

The use of DROs was extremely helpful for both of these purposes and especially the latter. Because the size, shape, and intensity values of each DRO was known, each object provided a unique opportunity to identify when software packages were calculating different values for the features of interest and allowed investigators opportunities to understand the underlying causes behind these differences. Some of these will be discussed in the following paragraphs.

The use of specific lesions in patient data sets was also very helpful in pursuing this project's primary purpose, in that the lesions studied did represent clinical objects of interest. In addition, the 10 cases selected represented different challenges in

terms of lesion size and composition, slice thickness, and slice spacing as well as manufacturer (which influences expected aspects, such as contrast-to-noise ratios, as well as unexpected ones, such as the DICOM field rescale intercept).

When our investigators performed this task under very tightly controlled conditions of the same feature definitions, using same image data and the same VOI definition, some features showed excellent agreement (CV $<$ 1%). This was true for the approximate volume, mean intensity, standard deviation, and kurtosis features. The definitions of these features are relatively straightforward and there are not many choices in parameters when calculating these features. The one exception was in the kurtosis feature where some software packages do not include the Fisher adjustment (subtracting 3 from the sum). With the use of the DROs, this was readily identified and corrected.

For other features, the agreement between software packages was not quite as good. The 3D diameter feature did show excellent agreement between software packages for the DROs (CV $<$ 1%), but showed slightly larger variation when more irregularly shaped objects were used in the patient data sets; however, the resulting variability was still reasonably low (CV $<$ 3%) across all 10 patient data sets and software packages. The 2D diameter feature (which was not actually defined by the IBSI) showed larger variation than the 3D diameter. This increased variation may be the result of differences in approach that were allowed for this study. As described above, some sites calculated diameter on the slice with the largest area, while others calculated the diameter for each slice and used the maximum from all slices. Thus, while the DRO results showed little variation (CV $<$ 1% for the spherical phantoms and CV $<$ 5% for the shape varying phantom), there were larger variations (CV = 10%) for the patient data sets, which are likely because of the variations in approach described here. Our investigation did not constrain which approach should be used (nor did we specify the image on which the diameter calculation should be made) so as to reflect conditions typically encountered in the clinic (such as making RECIST measurements for a clinical trial).

The surface area and sphericity features showed larger variability (CV = 12%–18%) in both DRO and patient data set results. First, it should be noted that these 2 features are related, in that sphericity uses the surface area (and volume) in its definition. It should also be noted that the IBSI description of the surface area feature is specific to the use of the mesh-based calculation method. This does require some specification (either implicitly or explicitly) of the size or number of triangles used in the mesh, which was not constrained in this investigation. In addition, some of the software packages used may not have used the mesh-based approach, but may have calculated an approximation to surface area by counting the areas of surface voxels. This disparity in approaches is likely one source of variability in this study.

Finally, the GLCM entropy feature showed the largest amount of variability across software packages. For the DROs, the CV was 51% for the intensity varying phantom and much higher values for the uniform and shape varying phantoms. These latter high values were explored and found to be the product of 2 software packages that calculated nonzero values for a completely uniform object. This was investigated and determined that when forming the co-occurrence matrices, some software packages check to determine if both the source and destination voxels are contained within the VOI while others only check to determine if the source voxel is contained within the VOI. In the former situation, both voxels are within the VOI; for the DRO with uniform voxel values, there will be no intensity variation, and the GLCM entropy value will be 0. In the latter situation, near the boundary of the VOI, the destination voxel may be outside the VOI and then voxels outside the VOI are included in the calculation. For the DRO with uniform values (100 HU) inside the boundary and uniform but different values (−1000 HU) outside the boundary, this will lead to nonzero values for the GLCM entropy values.

Furthermore, for the patient image data sets, the GLCM entropy values showed larger variations (CV = 36% across 10 cases) than other features. We hypothesized that this was due to the large number of choices and parameter settings that go into both the formation of the GLCM matrix and the calculation of the feature itself. These include (but are not limited to) issues related to: (1) any interpolation scheme used either for the image data at the beginning of the matrix formation or later during the formation process; (2) the choice of distance and angle or the use of multiple distances and angles (and any subsequent averaging over distances and/or angles); (3) the discretization scheme (scale and number of quantization bins) used; and (4) the feature aggregation scheme (IBSI describes 6 different schemes).

To determine if the effects of these choices and parameter settings could be mitigated, investigators agreed to a reference set of conditions described in the section on Harmonization of GLCM entropy and recalculated their results. Our results showed that the variability was reduced (CV reduced from 36% to 19%), but still did not approach that of the features with excellent agreement. This indicates that even this level of harmonization was not sufficient to achieve excellent agreement across software packages with this complex feature.

Also, even after harmonization, the results in Table 4 indicated some possible dependence of agreement (CV%) on lesion size. Specifically, lesions 3 and 8 were the smallest lesions, but showed larger than average CV values. In addition, lesion 4 was a calcified lesion and it also showed large CV values across packages. Thus, lesion size and density may have an impact on this specific feature value, but we did not have a large enough sample size of lesions to explore this fully in the current study.

In summary, this study has shown that excellent agreement can be achieved for some features when standardized definitions of features are provided such as those in the IBSI reference manual. This led to excellent agreement among software packages for features such as approximate volume, mean intensity, standard deviation, kurtosis (with Fisher adjustment), and even 3D diameter.

For more complex features such as surface area, sphericity, 2D diameter, and even GLCM entropy, very complete definitions of specific approaches and parameter settings used (eg mesh size, interpolation schemes, quantization levels, etc.) are needed to achieve the levels of agreement observed in the other features.

There are several limitations of this study. This study does not address all sources of variation in the calculation of radiomic features. For example, this study did not address issues related to image segmentation and the variability introduced when different VOIs are identified for a given object (7) or the effects of

different image acquisition and reconstruction conditions on feature variability (16, 17). This study also does not address the very important question of which features actually provide information relating to a specific clinical task; for example, this study did not address which features are useful in discriminating benign from malignant lesions or responders from nonresponders in clinical trials. As noted previously, some features may be shown to be very stable over a wide range of segmentation or acquisition and reconstruction conditions (eg feature value may be constant), but may not necessarily be useful in performing a given clinical task such as differentiating between benign and malignant lesions. This investigation was also limited to individual radiomic feature values and did not extend to investigations into "feature signatures" or functions that use multiple radiomic features, which would of course depend on the calculation of these underlying features. While evaluating the robustness of radiomic signatures was beyond the scope of the current investigation, this would be a very important next step in evaluating the use of combinations of radiomic features in clinical tasks.

In addition, this study investigated a limited number of features that represent a small fraction of the radiomic features described in the literature and specifically in the IBSI reference manual (13). This was done intentionally to keep the study size manageable and to allow thorough investigation of the features that were included in the study. Therefore, the definition and implementation issues for all radiomic features were not all addressed in this study; however, we believe that this exercise is instructive in identifying some underlying issues that need to be addressed with regard to the reproducibility of radiomic features.

It should be noted that 3 different sites all submitted results using the open source software package PyRadiomics (10, 23). This may have led to somewhat artificially low CV (CV%) results in some instances. This concern is somewhat offset by the observation that several features were observed to have very low (CV $<$ 1%) across all software packages. On the other hand, even these 3 sites did not achieve complete agreement for the GLCM entropy feature, despite the harmonization steps taken, although the agreement was better than that across all packages (mean CV was 10% for just PyRadiomics, compared to 19% observed across all packages).

The sites that used PyRadiomics noted that the package is built nightly, which may result in slightly different versions of the code and a potential source of variation. Moreover, because PyRadiomics can use both DSO and NIfTI segmentation objects (and 2 of our sites used DSOs while 1 used NIfTI), this may contribute to slight variations in feature values.

In addition, there were several lessons learned by the participating sites in this study. Many of these relate to assumptions that sites (and specific software packages) have made that are often implicit or are the result of implementation of certain mathematical operations or parameters for calculation of a feature by a site or in a specific package. These variations are difficult to identify without going through an exercise such as using a reference set of images (both DROs and patient data sets) and comparing results. These lessons include the following:

1. When calculating volume, slice spacing should be used rather than slice thickness [DICOM field (0018, 0050)]. Slice spacing is not a specific DICOM field, but can be calculated as the difference in either the z location of adjacent images using (0020, 0032) or (0020, 1041). This can be used to calculate features such as approximate volume. Using slice thickness can lead to overestimation of volume values in some cases. In many CT studies, the slice spacing and slice thickness values may be the same (eg, 2.5-mm-thick images spaced 2.5 mm apart), but the online supplemental Appendix shows that 6 of our 10 patient cases had slice spacing values different from the slice thickness. In these cases, using slice thickness instead of spacing can lead to erroneous volume values.
2. Identification of when interpolation of image data is taking place. In some cases, this interpolation takes place explicitly when image data are read in, and in other cases, it may occur implicitly or as part of the calculation of a feature. The interpolation as well as the interpolation scheme (eg, nearest neighbor, trilinear, tricubic convolution, tricubic spline interpolation, etc.) can lead to variations in feature values.
3. The methods for interpretation of VOI boundaries and the management of "holes" within a VOI (eg whether the software would "fill the hole" or leave it "as is") could be potential sources of variance between software packages.
4. There were also several feature-specific issues. For example, when forming GLCM matrix, there were differences between software packages in whether both the source and destination voxels are checked to be within VOI or if just the source voxel was checked. Another example was that the surface area feature is defined by a mesh-based representation in the IBSI reference manual (Section 3.1.3 of that manual), but our sites reporting feature values may not use that same approach. In addition, even with the mesh-based representation, there still is variability introduced by the selection of parameters such as the mesh size.
5. Finally, the GLCM entropy feature was specifically chosen for analysis in this study to be representative of the complexity of commonly used texture features in radiomics studies. This feature has several parameters and approaches involved in its calculation (as identified above), and the choice of each can contribute to variability. Our results showed that some issues could be mitigated by using similar parameters, but others may require re-coding of the software. In general, such complex features can be replicated universally only if a common software package (eg, PyRadiomics or similar) is used. However, if such an approach is considered, the community should thoroughly evaluate the specific definitions and implementations of the features in the potential common package and reach consensus that it is acceptable as the "gold standard" before recommending its use.

Based on the initial purpose of this study, its results and the lessons learned from this study, this group recommends the following to improve the reproducibility of results:

1. That authors provide detailed description of the image analysis, image pre-processing, and feature definitions in

future studies reporting results involving radiomic features so that an independent site could reproduce the results of the study.
2. That IBSI feature definitions be used in manuscripts wherever possible, including the feature coding descriptors whenever possible.
3. That DICOM develop a DICOM standard for reporting features according to the definitions and features described in the IBSI reference manual.
4. That reporting structures support flexibility that will allow additional information regarding feature descriptions (in case the IBSI definition is not quite complete) and will allow for further development of future features to be included.
5. That publication of results should include both the software name and version number. Although this might be difficult for open source software packages that undergo frequent builds and releases, we encourage the description of the used version in as much detail as possible.
6. That radiomics software packages follow unique release version numbering.
7. That software source code be released whenever possible (as has been done by several packages used in this study—see Table 1). Mechanisms such as GitHub have proven to be extremely useful in this regard.

CONCLUSIONS

From this study, we are able to conclude that some radiomic features (approximate volume, mean intensity, standard deviation, kurtosis with Fisher adjustment, and 3D diameter) can be calculated across different software packages and achieve high levels of agreement. However, we were not able to show the same level of agreement for other features such as surface area, sphericity, 2D diameter, and GLCM entropy. For these features we observed larger variability either due to variations in user selections (2D diameter), implementation approach, or parameter selections. This indicates that even when a reference set of radiomic feature definitions is used, additional information may be required to reproduce results obtained with different software packages. Thus, efforts to further define options in how radiomic features are calculated as well as encouragement in reporting these details can help improve one aspect of radiomic feature variability and ultimately contribute to the reproducibility of these features as they are introduced to clinical trials and ultimately to clinical practice.

Supplemental Materials

Supplemental Appendix: https://doi.org/10.18383/j.tom.2019.00031.sup.01

ACKNOWLEDGMENTS

Equal contribution: M.M.G. and S.N. contributed equally to this work and should be considered co-first authors.

We gratefully acknowledge the following sources of support: David Geffen School of Medicine at UCLA—U01CA181156; Stanford University School of Medicine—U01CA187947 and U24CA180927; University of Michigan—U01CA232931; University of Washington—R50CA211270 and U01CA148131; University of South Florida—U24CA180927 and U01CA200464; Moffitt Cancer Center—U01CA143062, U01CA200464, and P30CA076292; UC San Francisco—U01CA225427; BC Cancer Research Centre—NSERC Discovery Grant: RGPIN-2019-06467; Columbia University—U01CA225431; Center for Biomedical Image Computing and Analytics at the University of Pennsylvania—U24CA189523 and R01NS042645; Massachusetts General Hospital—U01CA154601 and U24CA180927.

The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

REFERENCES

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–446.
2. Xing L, Li R, Napel S, Rubin D. Radiomics and Radiogenomics: Technical Basis and Clinical Applications. Boca Raton, FL: CRC Press, 2019.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278:563–577.
4. Napel S, Mu W, Jardim-Perassi BV, Aerts H, Gillies RJ. Quantitative imaging of cancer in the postgenomic era: radio(geno)mics, deep learning, and habitats. Cancer. 2018;124:4633–4649.
5. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
6. Clarke LP, Nordstrom RJ, Zhang H, Tandon P, Zhang Y, Redmond G, Farahani K, Kelloff G, Henderson L, Shankar L, Deye J, Capala J, Jacobs P. The Quantitative Imaging Network: NCI's historical perspective and planned goals. Transl Oncol. 2014;7:1–4.
7. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P, Sieren JC, Uthoff J, Dilger SK, Driscoll B, Yeung I, Hadjiiski L, Cha K, Balagurunathan Y, Gillies R, Goldgof D. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography. 2016;2:430–437.
8. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A. Shinohara RT. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. 2019. https://arxiv.org/abs/1811.02629.
9. Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative Image Feature Engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. J Digit Imaging. 2018;31:403–414.
10. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts H. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104–e107.
11. Davatzikos C, Rathore S, Bakas S, Pati S, Bergman M, Kalarot R, Sridharan P, Gastounioti A, Jahani N, Cohen E, Akbari H, Tunc B, Doshi J, Parker D, Hsieh M, Sotiras A, Li H, Ou Y, Doot RK, Bilello M, Fan Y, Shinohara RT, Yushkevich P, Verma R, Kontos D. Cancer imaging phenomics toolkit : quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. J Med Imag. 2018;5:011018–1–011018–21.
12. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. An open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42:1341–1353.
13. Zwanenburg A, Leger S, Vallieres M, Lock S. The image biomarker standardisation initiative. (Online). Available: https://arxiv.org/abs/1612.07003. (Accessed 20-December-2019).
14. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Wang X, Yang H, Tan Y, Gillies R, Napel S. A Comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. J Digit Imaging. 2016;29:476–487.
15. Young S, Kim HJ, Ko MM, Ko WW, Flores C, McNitt-Gray MF. Variability in CT lung-nodule volumetry: effects of dose reduction and reconstruction methods. Med Phys. 2015;42:2679–2689.
16. Lo P, Young S, Kim HJ, Brown MS, McNitt-Gray MF. Variability in CT lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. Med Phys. 2016;43:4854.

17. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, Schwartz LH. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428.
18. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, van Beek EJR, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DP-Y, Roberts RY, Smith AR, Starkey A, Batra P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Vande Casteele A, Gupte S, Sallam M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY, Clarke LP. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38:915–931.
19. American College of Radiology, "Lung CT Screening Reporting & Data System (Lung-RADS)." (Online). Available: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads. (Accessed: 20-December-2019).
20. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst, Man, Cybern. 1973;SMC-3:610–621.
21. Jaggi A, Mattonen SA, McNitt-Gray M, Napel SN. Stanford DRO Toolkit: digital reference objects for standardization of radiomic features. Tomography. 2020;6:111–117.
22. The Cancer Imaging Archive (TCIA). (Online). Available: cancerimagingarchive.net. (Accessed: 20-December-2018).
23. Py Radiomics from the Computational Imaging & Bioinformatics Lab - Harvard Medical School. (Online). Available: https://www.radiomics.io/pyradiomics.html. (Accessed: 20-December-2019).
24. Echegaray S, Bakr S, Rubin D, Mattonen SA, Napel S. Quantitative Image Feature Engine Github page. (Online). Available: https://github.com/riipl/3d_qifp. (Accessed: 16-December-2019).
25. Brown MS, McNitt-Gray MF, Pais R, Shah SK, Qing P, Da Costa I, Aberle DR, Goldin JG. CAD in clinical trials: current role and architectural requirements. Comput Med Imaging Graph. 2007;31:332–337.
26. Brown MS, Shah SK, Pais RC, Lee YZ, McNitt-Gray MF, Goldin JG, Cardenas AF, Aberle DR. Database design and implementation for quantitative image analysis research. IEEE Trans Inform Technol Biomed. 2005;9:99–108.
27. Brown MS, Lo P, Goldin JG, Barnoy E, Kim GH, McNitt-Gray MF, Aberle DR. Toward clinically usable CAD for lung cancer screening with computed tomography. Eur Radiol. 2014;24:2719–2728.
28. PMOD Technologies LLC, "PMOD Software." (Online). Available: https://www.pmod.com/web/. (Accessed: 20-December-2019)
29. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging. 2017;44:151–165.
30. Chan HP, Wei D, Helvie MA, Sahiner B, Adler DD, Goodsitt MM, Petrick N. Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space. Phys Med Biol. 1995;40:857–876.
31. Cha K, Hadjiiski LM, Chan Hp, Samala Rk, Zhou C, WJ. MiViewer Interface for computer-aided decision support and observer performance evaluation of bladder cancer treatment response assessment. Demonstration at the live demonstration workshop, SPIE International Symposium of Medical Imaging, Orlando, FL. February 11-16, 2017. (http://spie.org/Documents/ConferencesExhibitions/MI/SPIE2017-LiveDemo-list-2017.pdf)
32. Ashrafinia S. Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics. PhD dissertation. Baltimore, MD, 2019. Available at: https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/61551/ASHRAFINIA-DISSERTATION-2019.pdf?sequence=1&isAllowed=y (Accessed: 20-December-2019).