Original Article, PET/CT.

Radiomics Feature Selection in ^{18}F-FDG PET Imaging: Investigation of the Most robust and Reproducible Candidates.

Khalil, MM[1].

Department of Physics, Faculty of Science, Helwan University, Egypt.

ABSTRACT

PET radiomics can reveal a lot of clinical information but with many technical and physical variables to be tackled. The aim of this study was to find out the most reproducible and robust PET radiomics that can be used as benchmark for future clinical studies. **Materials and Methods**: Two phantom datasets were retrieved from the large Cancer Imaging Archive repository (NIH, USA) and employed for further radiomics feature extraction (i.e. 108 features) and analysis including calculation of coefficient of variation (CV) and percent deviation. Further testing was carried out to investigate effect of sphere size and contrast level as well as imaging scanner on feature reproducibility. Feature correlation with tumor TNM stage was also performed using Spearman correlation and Kendall tau statistical tests. Principle component analysis was used to reduce the large feature number to principle components using the eigen vector and eigen value. **Results:** The (CV) and percent deviation revealed 56 features out of 108 that have less than or equal to 10% variability. When features were compared in terms of high and low image contrast, there was 46 features that showed less sensitivity to different concentrations. Inter-scanner variability testing has reduced the number of features to 36 out of the 46 features. Analysis of PCA results showed that 4 components can account for 90.3%, namely, SUV max, inverse difference moment normalized; size zone non-uniformity normalized, and short run low gray level emphasis. When correlating the 108 features with patient status, two features only showed significant correlation with tumor stage namely maximal correlation coefficient and flatness.

Conclusions: The 108 features were reduced to a lower percentage of features while maintaining large percentage of data variance and feature reproducibility. The combination of texture feature of robust technical qualifications to those of clinical value would finally improve the clinical decision model.

Key words: *PET/CT Radiomics, Texture Features, Reproducibility & Dimensionality Reduction.*

Corresponding Author: Khalil, M. **E-mail:** magdy_khalil@hotmail.com.

INTRODUCTION:

Cancer remains one of the most influential disease on different aspects of human life [(1)]. Imaging cancer biomarkers have been increasing in the last two decades due to recent advances in imaging sciences and technologies [(2, 3)]. Tumor heterogeneity has been recently identified as one important aspect that reflects the degree of tumor aggressiveness, drug resistance and plays an important role in patient management [(4)]. Nuclear medicine provides unique opportunities in diagnosing patient with several malignancies using F18-Flurodeoxylucose (F18-FDG) and positron emission tomography combined with computed tomography (PET/CT) imaging systems. The recent advances in data analysis, modeling algorithms and computational tools have permitted the introduction of a new term called "radiomics" and "radiogenomics" analogous to new developments in genomics, proteomics, transcriptomic and other "omics" sciences.

Radiomics is an emerging computational field that deals with extracting statistical and physical features of the images beyond physician perceptive and recognition capabilities [(5)]. There are wide range of features that researchers and scientists were able to derive from different imaging modalities including, PET, CT and magnetic resonance imaging data. However, the scientific community has recently released a standardized framework that enables uniform guidelines to be followed called "Image biomarker standardization initiative, IBSI" providing definitions, benchmark data and values as well as calculation methodologies for high throughput imaging biomarkers extraction [(6)]. Several studies have been conducted to select those features of high stability, reproducibility and reliability.

At the top of that, the pathological/clinical correlation of these features with patient diagnosis, prognosis as well as stratification must be defined and properly addressed.

On the technical or physical side, there several parameters and variables that hindered the prompt implementation of radiomics into clinical routine including variation of acquisition, reconstruction, and image analysis in addition to variables associated with inter-scanner and system specific characteristics [7,8]. Moreover, vendor specific parameters add more complexity to this process which become even more with end-user specific preferences. In multicenter clinical trials, all the physical and technical issues must be harmonized in order to reduce variation among the contributing clinical centers.

When the physical and technical variables are addressed, then the modeling algorithm need to be carefully designed through robust clinical studies. A number of reports have investigated the utility of radiomics with potential results in application to patient diagnosis and prognosis with documented varieties of patient survival and risk stratifications [9,10].

Despite of the large records of publication and literature on radiomics data analysis, there is no consensuses on which of those features are the most reproducible and repeatable with direct use in the clinic.

However, there are few studies that dealt with radiomics feature variability and accuracy in phantom studies.

In a recent systematic report, there were 6 phantom radiologic studies that comprised 5 CT and only 1 phantom PET study [11]. Moreover, no observed pattern was detected for PET texture features among the several studies included in the analysis. Therefore, the aim of this study was to search for the most appropriate features of PET data that provide less variation and deviation as well those features that are clinically relevant through standard phantom datasets.

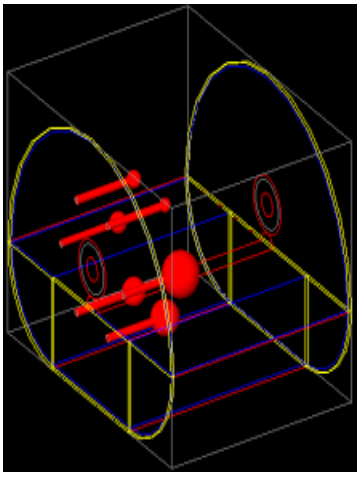
MATERIALS AND METHODS:

Datasets: Quantitative Imaging network (QIN) PET Phantom Collection contains PET phantom scans originally utilized by the Quantitative Imaging Network PET Segmentation Challenge to assess the variability of segmentation methods and image quantitative analysis results [12]. These images were acquired using NEMA IEC Body Phantom Set Model PET/IEC-BODY/P and thus the ground truth of sphere volume and injected activity are known and defined in prior.

The phantom was scanned at 2 medical institutions using different imaging systems namely Biograph (Siemens Medical Solutions) and Discovery STE (GE Healthcare, Inc.) in University of Iowa and University of Michigan respectively. The source files are called "TCIA_QIN_PET" [13]. **Phantoms:** The NEMA IEC PET phantom provides several opportunities for measuring the robustness of the radiomics feature. The size variation of the sphere volumes is one important aspect such that one can look at the degree of correlation of different image features with lesion volume/size. The repeated measurements (here was 10 identical measures) provides also an opportunity to elaborate on the ensemble variance and how this could impact feature stability and/or reproducibility. The list mode data acquisition was also helpful to reconstruct the whole data set into one time points of 30 min providing reconstructed data of high statistical quality with minimal noise and being reference for other low statistical data measures. Therefore, the test-retest data analysis can be measured for individual feature metrics. ***Table (1)*** describes the physical phantom used in data acquisition. It is worth noting that the spheres have various dimensions and size including spherical and ellipsoidal geometry in both axial and horizontal directions. In high and low contrast scans, the phantom was filled up with an activity ratio of almost 10:1 and 4:1 respectively.

Table (1): Diagrammatic representation and phantom characteristic used in data acquisition. It should be noticed that the same phantom was used in the two institutions.

	Shape and orientation	Diameter/size (mm)	Volume (ml)
Sp1	Sphere	28	11.49
Sp2	Ellipsoid, axial	34x17x17	5.14
Sp3	Ellipsoid, axial	26x13x13	2.30
Sp4	Ellipsoid, horizontal	26x13x13	2.30
Sp5	Sphere	13	1.15
Sp6	Ellipsoid, Horizontal	20x10x10	1.04



"Sp" stands for sphere

Acquisition: Data were identically acquired with the two different contrasts each with 30 min total time using list mode acquisition. The 30 min data acquisition was then divided into 10 scans each with 3 min to provide a clinically relevant data acquisition. And the 30 min acquisition provided a more stable and high statistical certainty in quantitative measurements while the later 3 min sequential acquisitions were used to mimic clinical scans with equivalent noise and imaging characteristics and also offered an opportunity to investigate the associated scan reproducibility.

Image Reconstructions: Images were reconstructed using diverse reconstruction parameters to create PET data of challenging properties versus radiomics feature extraction and analysis. Raw data revealed from the Siemens biograph was reconstructed using 7 mm Gaussian post-filter while the other data produced by GE Discovery was reconstructed with a sharper cut-off value of 3 mm using the same filter. In both reconstructions, the iterative ordered-subset expectation maximization (OSEM) was used [(14)]. The former provides images with high smoothing characteristics while the later filtering was applied to give images with high spatial resolution. This is again to increase data diversity providing an opportunity to derive the most robust features that able to maintain stability and data reproducibility.

Standardized uptake Value (SUV): SUV max and SUV mean have been extensively used in literature for patient diagnosis, prognosis and response monitoring. The acquired TCIA_QIN_PET data provided several opportunities to investigate the impact of different scanning systems with different noise texture and resolution capabilities (Discovery STE vs. Biograph) of different manufacturers, inter-scan repeatability (i.e. 10 scans were acquired each with 3 min), reconstruction parameters (3 mm vs. 7 mm FWHM, Gaussian filter), as well as providing reference images of high statistical quality (30 min data acquisition) at high and low contrasts. The results of the SUV max and SUV mean were reported to confirm the consistency of the values with the injected radioactivity.

Sphere/Lesion segmentation: Lesion or sphere segmentation was performed using a semi-automated approach that transform the segmentation process into graph based optimization problem recently reported [(15)]. The PET segmentation extension plugin was used to segment the 6 different spheres of each acquisition of the two clinical sites.

The process starts by locating the center of the sphere and the algorithm then determines the sphere contours through a built-in cost function that provide lesion surface. It has also the capability to encounter situations when lesion has numerous heterogeneity as well as features to refine the lesion contours whenever necessary [15].

Feature Extraction: The 3D slicer software package was used in sphere segmentation to calculate the various features of the PET data described earlier. The following extension/plugins were used in importing the data sets: dicom-plugin, dicom scalar volume, multivolume importer, dicom longitudinal PET/CT, dicom PET SUV and dicom slicer data bundle plugins. These modules enabled to obtain the appropriate measures of SUV as exactly measured and calculated on the original scanners and processing workstations. Morphological, shape, first and higher order statistical features were calculated using the pyradiomics modules installed as extension plugin in the 3D slicer version 4.10.2.

Since these matrices are calculated based on voxel intensity, location, neighborhood and spatial arrangement with other voxels, let us define some variables used in various feature formula: Let X be a set of Np voxels of the region of interest under investigation, P(i) defined as the histogram of the first order matrix with Ng being defined as discrete intensity. The latter is the number of bins which have non-zero value and limited by the bin width preset in the 3D slicer software package. The bin number selected was 0.5 as recommended in previous reports. P(i) is the normalized first order histogram and defined as P(i)Np. Here is a summary of some examples taken from every radiomics matrix used in features extraction.

First order features: This type of features is concerned with voxel count distribution in each region of interest. An example of first order statistics is the energy and defined as

$$Energy = \sum_{i=1}^{Np} X(i) + c$$

Where c is an optional parameter to shift values in X to positive values. Another important feature is entropy and defined as

$$Entropy = - \sum_{i=1}^{Ng} P(i) log_2(p(i) + \epsilon)$$

Other features are described in table 2 and their exact mathematical expressions are outlined in reference [5,6].

Shape features: Shape features are special set that mainly describe the ROI shape characteristics including sphericity, eccentricity, diameter and other geometric conformations that tumor mass may have. One example of the shape feature set is the volumetric measure and defined as

$$V_i = \frac{O_{ai}(O_{bi\,x}\,O_{ci})}{6}$$

Where Vi is the mesh volume and calculated from the triangle of the region of interest such that the face *i* (defined by the points a_i, b_i an c_i) of the tetra-hedron circumvented by that face and image origin is computed [16].

Gray Level Dependence Matrix (GLDM) Features (GLDM): This matrix finds out the relationship between a given voxel with gray level intensity *i* with its neighborhood voxel j to construct the gray level dependence matrix P(i,j). One example of GLDM is the Small Dependence Emphasis (SDE) defined as

$$\text{SDE} = \frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Na}\frac{P(i,j)}{i^2}}{N_z}$$

Greater emphasized values of SDE reflect small dependences and textures of low homogeneity.

Gray Level Co-occurrence Matrix (GLCM): This matrix reflects how frequently two voxels with specific values and within specific spatial correlation are repeated within the region of the interest to build what is called gray level co-occurrence matrix. Autocorrelation is one important example and defined as

$$Autocorrelation = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng} P(i,j)ij$$

Autocorrelation reflects fineness and coarseness of the lesion

Gray Level Run Length Matrix (GLRLM): The GLRLM matrix focuses on the length over which the voxel intensity is frequently repeated. In constructing the P(*i*,*j*|φ) matrix, the element (*i*,*j*) refers to the number of segments with gray intensity that takes place in the region of interest along the angle φ (phi). Short run emphasis belongs to GLRLM matrix and defined as

$$SRE = \frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr}\frac{P(i,j)}{j^2}}{Nr_{\varphi}}$$

When the lesion contains large frequency of short repeated gray intensity levels, the SRE is increased reflecting fine textures of the tumor.

Gray Level Size Zone Matrix (GLSZM): This is an additional matrix that reflects how small a zone of the region of interest is frequently occurred such that the constituent voxels have the same gray intensity. Gray level Non-Uniformity is one form that belongs to the GLSZM and defined as

$$GRN = \frac{\sum_{i=1}^{Ng}(\sum_{j=1}^{Ns} P(i,j)^2}{Nz}$$

The GRN with lower magnitude is indicative of more tumor homogeneity.

Neighbouring Gray Tone Difference Matrix (NGTDM): The NGTDM measures the difference between a given voxel and the neighboring voxel within a tolerance distance. Coarseness is one famous example derived from the NGTDM matric and mathematically described as

$$Coarsness = \frac{1}{\sum_{i=1}^{Ng} P_i S_i}$$

Coarseness is a measure of intensity change over neighboring voxels and greater values reflects localized tumor uniformity.

Radiomics module in the 3D slicer was then used to extract the features of each sphere and tabulated for further analysis [17].

The images of high-count statistics were used as references for images acquired with short time intervals (i.e. 3 min). The 10 reframed 3 min reconstructed images were used to measure the reproducibility of features under normal and clinically relevant noise conditions. A number of 108 features were extracted from each sphere including high and low count statistics as acquired using GE discovery STE and Siemens Biograph.

Evaluation metrics: To select those features of low sensitivity to several acquisitions and reconstruction parameters, the coefficient of variation was computed to measure the degree all features described in table 1. The deviation of the 10 measurements from the reference 30 min acquisition was also considered for selecting those feature that would provide less percent deviation. The texture features that achieved coefficient of variation and percent deviation equal to or less than 10% and 5% were reported and used for further analysis. Since the geometric features are not so sensitive to repeated acquisitions and count statistics, they were removed from the coefficient and percent deviation computation and subsequent analysis. The coefficient of variation was calculated using the following formula:

$$CV = standard\ deviation(SD)/mean$$

Table (2): The radiomics texture features used in data analysis. The major shape, first-order as well as higher order matrices are described with the associated features and their number.

Shape	Voxel Volume (VV), Maximum 3D Diameter (M3D), Mesh Volume (MV), Major Axis Length (MAL), Sphericity, Least Axis Length (LAL), Elongation, Surface Volume Ratio (SVR), Maximum 2D Diameter (M2D), Slice Flatness, Surface Area, Minor Axis Length (MAL), and Maximum 2D Diameter Row (M2DDR) and Maximum2D Diameter Column (M2DDC)	14
First-order	Interquartile Range (IR), Skewness, Uniformity, Median, Energy, Robust Mean Absolute Deviation (RMAD), Mean Absolute Deviation (MAD), Total Energy (TE), Maximum, Root Mean Squared (RMS), 90Percentile, Minimum, Entropy, Range, Variance, Standard Deviation (SD), 10Percentile, Kurtosis, Mean.	19
GLDM	Gray Level Variance (GLV), High Gray Level Emphasis (HGLE), Dependence Entropy, Dependence Non-Uniformity (DNU), Gray Level Non Uniformity (GLNU), Small Dependence Emphasis (SDE), Small Dependence High Gray Level Emphasis (SDHGLE), Dependence Non-Uniformity Normalized (DNUN), Large Dependence Emphasis (LDE), Large Dependence Low Gray Level Emphasis (LDLGLE), Dependence Variance (DV), Large Dependence High Gray Level Emphasis (LDHGLE), Small Dependence Low Gray Level Emphasis (SDLGLE), Low Gray Level Emphasis (LGLE)	14
GLCM	Joint Average, Sum Average, Joint Entropy, Cluster Shade, Maximum Probability, Idmn, Joint Energy, Contrast, Difference Entropy, Inverse Variance, Difference Variance, Idn, Idm, Correlation, Autocorrelation, Sum Entropy, MCC, Sum Squares, Cluster Prominence, Imc2, Imc1, Difference Average, Id, Cluster Tendency.	24
GLRLM	Short Run Low Gray Level Emphasis (SRLGLE), Gray Level Variance (GLV), Low Gray Level Run Emphasis (LGLRE), Gray Level Non Uniformity Normalized (GLNUN), Run Variance, Gray Level Non Uniformity (GLNU), Long Run Emphasis (LRE), Short Run High Gray Level Emphasis (SRHGLE), Run Length NonUniformity (RLN), Short Run Emphasis (SRE), Long Run High Gray Level Emphasis (LRHGLE), Run Percentage, Long Run Low Gray Level Emphasis (LRLGLE), Run Entropy, High Gray Level Run Emphasis (HGLRE), Run Length Non Uniformity Normalized (RLNUN).	16
GLSZM	Gray Level Variance (GLV), Zone Variance, Gray Level Non Uniformity Normalized (GLNUN), Size Zone Non Uniformity Normalized (SZNUN), Size Zone Non Uniformity (SZNU), Gray Level Non Uniformity (GLNU), Large Area Emphasis (LAE), Small Area High Gray Level Emphasis (SAHGLE), Zone Percentage, Large Area Low Gray Level Emphasis (LALGLE), Large Area High Gray Level Emphasis (LAHGLE), High Gray Level Zone Emphasis (HGLZE), Small Area Emphasis (SAE), Low Gray Level Zone Emphasis (LGLZE), Zone Entropy, Small Area Low Gray Level Emphasis (SALGLE)	16
NGTDM	Coarseness, Complexity, Strength, Contrast, Busyness	5
Total Number of Features		108

Where mean and standard deviations were taken for each sphere across the 10 scans. The percent deviation (%) was measured using the median of the 10 sequential measurements versus the images acquired with high statistical counts multiplied by 100. The use of median was intentional as the data showed a non-normal distribution. Similarly, texture features that evaluated CV and percent deviation less than 5% and 10% were filtered and tabulated for comparison.

Principle Component analysis (PCA):

One of the most commonly but potentially used data reduction techniques is principle component analysis (PCA) in which the algorithm creates new components that account for overall variation in given data set such that features/variables within each component are linearly combined and correlated.

Furthermore, the other components contain variables that are not in correlation with those of the first component while able to explain more variance of the data. This mathematical technique enabled us to select lower number of features out of the 108 features contained in the original dataset. It mainly consists of deriving the eigen vector (principle component) and eigen value. The former is used to determine the direction of new feature space while the latter determines the magnitude.

Statistical Analysis:

Radiomics data analysis produces large amount of data and it is not necessary all features indicate significant tumor heterogeneity or correlation with clinical outcome. Therefore, attempts here were made first to look for those features which able to provide better reproducibility and stability using imaging data of the acquired phantom data. As highlighted in ***table 1***, the spheres have different geometry, shape and count distributions due to their relative size. We have used three different measures to search for features that can resist variations in counting statistics, variability introduced due to short time scanning, image filtering, and different scanning systems. The coefficient of variation, percent deviation and finally principle component analysis (PCA) were used to estimate those features that can account for the most variability while maintain as much as possible of diagnostic information.

The “prcop” function implemented in R statistical language was used along with “FactoMineR” and “FactoExtra” libraries to compute the PCA including eigen value, eigen vectors, feature as well sphere contribution to total variance [(18)].

Mann Whitney non-parametric test was used to comparing between the low and high contrast images acquired with high statistics using the two PET/CT imaging systems. A number of 13 patients who had head and neck cancer were retrieved from the cancer imaging archive database [(13)] and lesions were semi-automatically segmented using the above mentioned algorithm namely "just-enough interaction" graph based method [(15)]. There were 21 lesions extracted and tabulated for further analysis. To find out the relationship of the texture feature with tumor stage of the head and neck patients, Spearman correlation and Kendall tau were used to evaluate the strength and significance of correlation. Microsoft Excel version 16 and R statistical software package were used in data plotting and analysis

RESULTS:

The average SUV max and SUV mean for both PET/CT scanners including high and low contrasts for the 6 spheres are summarized in ***Table (3)***. The variation and percent deviation of the two SUV variants are also reported along with the SUV mean and SUV max measured for each sphere of the NEMA phantom. ***Table (4) and figure (2)*** describe those features that were able to achieve a measurement of CV lower 5% and 10%. The same was true for the percent deviation. This process resulted in smaller number of features. These selected features should have less variability in measurements and lower deviation from data acquired with high statistical certainty. They were then compared for each PET/CT scanner in terms of sphere contrast and also between the two imaging systems.

Effect of sphere contrast: ***Table (5)*** shows the Mann-Whitney results for comparing between the low and high contrast images acquired with high statistics using the two PET/CT imaging systems. A number of 22 features have shown a non-significant difference between the low and high contrast data acquired with Siemens Biograph whereas 24 features were obtained with GE Discovery STE. Therefore, a number of 46 features were found not to be affected by the change in image contrast in both imaging systems. Moreover, first order statistics have totally vanished in this test since they account for global estimate of tracer concentration. This could be clinically seen when patient received different range of doses, different imaging time after injection (early-delay imaging protocols), or different system sensitivity and/or imaging time.

Table (3): Results of SUV max and SUV mean which are the most commonly used quantitative metrics in routine practice of PET Imaging. Notice the gradual reduction of both values as the size decreases. This is consistent with partial volume effect phenomenon.

	UI				UW			
	Siemens Biograph				Discovery STE			
	High Contrast		Low Contrast		High Contrast		Low Contrast	
SUV max	SUV max	SUV mean	SUV max	SUV mean	SUV max	SUV mean	SUV max	SUV mean
Sp1	9.3	4.7	4.8	3.0	11.0	6.2	5.6	3.1
Sp2	8.0	3.7	3.7	2.5	10.5	5.3	5.2	2.7
Sp3	5.8	3.2	2.7	2.0	9.5	4.6	4.8	2.4
Sp4	6.5	3.3	3.1	2.1	9.8	4.4	4.7	2.3
Sp5	4.5	2.7	2.3	1.8	6.9	3.4	3.6	1.9
Sp6	4.8	2.8	2.5	1.9	9.4	4.2	4.9	2.2
	CV, median							
SUV max	17.55		11.17		9.3		15.36	
SUV mean	9.07		6.68		4.3		5.95	
	% Deviation							
SUV max	-34.47		-24.85		25.04		26.15	
SUV mean	20.14 ± 4.42		-15.79 ± 1.84		1.11 ± 0.45		1.22 ±0.21	

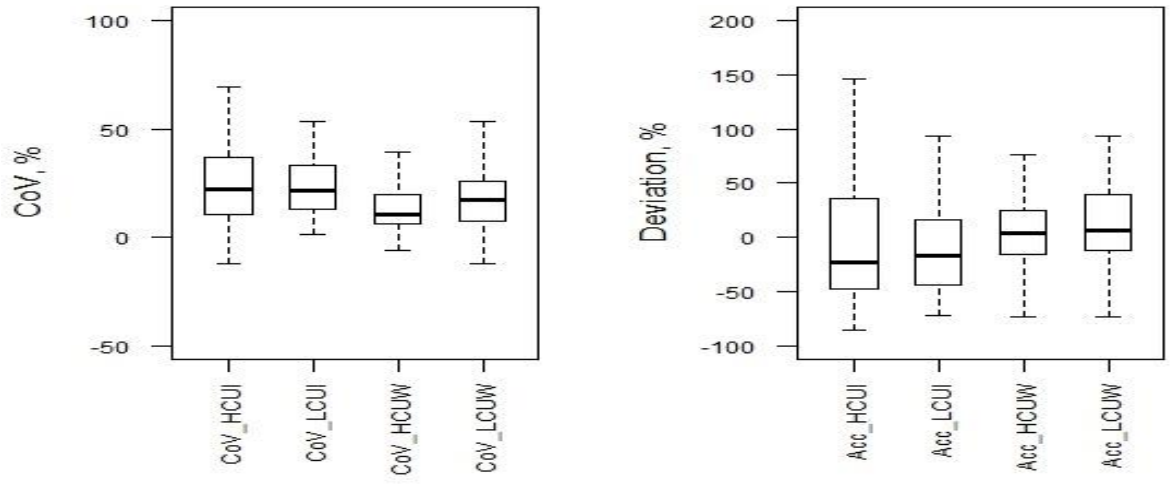


Figure 2: Distribution of coefficient of variation and percent deviations of all feature measured using the two different scanners at two different contrasts. The CV was measured for median of features over the 10 repeated acquisitions 2 min each. The % deviation was measured of how the median values of the 10 measurements are far from the images acquired with high count statistics.

Table (4): The texture features which achieved a coefficient of variation and percent deviation less than or equal to 5% and 10% derived from the 4 different data acquisitions. Data reported for individual measurement of CV and % deviation as well as their combination due to the reframed 10 datasets.

Texture feature with coefficient of variation less than or equal to 5%	**No.**
10Percentile, Dependence Entropy, Difference Entropy, Entropy, Idmn, Idn, Imc2, Joint Entropy, Kurtosis, Long Run Emphasis, Mean, Minimum, Root Mean Squared, Run Entropy, Run Percentage, Run Entropy, Run Length Non Uniformity, Run Length Non Uniformity Normalized, Run Percentage, Short Run Emphasis, Sum Entropy, Zone Entropy.	22
Texture feature with % deviations less than or equal to 5%	
90Percentile, Coarseness, Contrast, Correlation, Dependence Entropy, Dependence Non Uniformity, Dependence Non Uniformity Normalized, Dependence Variance, Idmn, Idn, Imc2, Interquartile Range, Joint Entropy, Kurtosis, Large Dependence High Gray Level Emphasis, MCC, Mean, Median, Minimum, Robust Mean Absolute Deviation, Root Mean Squared, Run Entropy, Run Length Non Uniformity Normalized, Run Percentage, Short Run Emphasis, Small Area Low Gray Level Emphasis, Sum Entropy, Zone Entropy.	28
Texture features with percent deviations less than or equal to 10% and coefficient of variation less than 10% [combination of CV and % deviation]	
Inverse Difference Moment Normalized (Idmn), Inverse Difference Normalized (Idn) Short Run Emphasis, Dependence Entropy.	4
Texture features with percent deviation less than 5% and coefficient of variation less than 5% [combination of CV and % deviation]	
Inverse Difference Moment Normalized (Idmn) and Inverse Difference Normalized (Idn)	2

Table (5): Mann Whitney nonparametric test to compare between low and high contrast imaging data sets for both clinical sites. These features have not shown a significant difference ($p>0.05$). A number of 22 features were found for UI (Siemens Biograph) and 24 features for UW (GE Discovery STE) out of the 56 filtered features presented for comparisons.

	UI (Siemens Biograph)	UW (GE Discovery STE)
Shape Feature	Voxel Volume, Maximum 3D Diameter, Mesh Volume, Major Axis Length, Sphericity, Least Axis Length, Elongation, Surface Volume Ratio, Maximum 2D Diameter Slice, Flatness, Surface Area, Minor Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row	Voxel Volume, Maximum 3D Diameter, Mesh Volume, Major Axis Length, Sphericity, Least Axis Length, Elongation, Surface Volume Ratio, Maximum 2D Diameter Slice, Flatness, Surface Area, Minor Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row
GLCM	Correlation	Correlation, Idmn, Idn
GLDM	Dependence Non-Uniformity. Large Dependence High Gray Level Emphasis. Small Dependence Low Gray Level Emphasis.	Dependence Entropy Dependence Non-Uniformity Large Dependence, High Gray Level Emphasis
GLRLM	Gray Level Non Uniformity. Run Length Non Uniformity.	Gray Level Non Uniformity. Run Length NonUniformity.
GLSZM	Small Area Low Gray Level Emphasis.	Small Area Low Gray Level Emphasis.
NGTDM	Coarseness.	Coarseness.

Effect of sphere size: The median CV among the 6 spheres was also computed for each scanner at the two different imaging contrasts and then plotted in figure 3. In this test, we have used the acquisition with large count statistics in both phantoms so that more statistical quality is maintained. The range of variation due to different sphere sizes used in data acquisition yielded quite large range of CV and percent deviation ***(Figure 3)***.

Inter-scanner comparison: ***Table (6)*** shows the texture features that demonstrated a non-significant difference when compared in terms of the scanner used in data acquisition. A significant number of shape features were found due to the fact that sphere geometry and dimension should not change when scanned with different scanners. The GLRLM matrix has shown also adequate number of features (i.e. n = 6) with insignificance to alternating the scanner model.

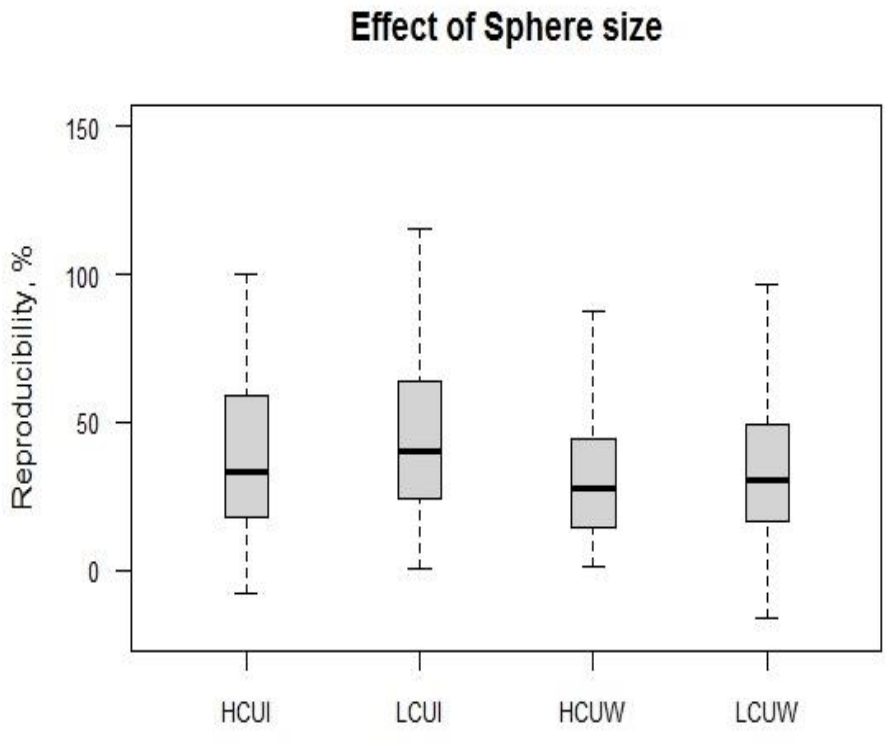


Figure (3): The reproducibility of the texture features due to changes in sphere/lesion size. It is obvious that the size remains an issue that need to be tackled due to the wide variation seen in all acquisition using the two PET/CT systems. The reproducibility is measured using the same formula for coefficient of variation.

Table (6): A summary of features that have not been affected by the type of the scanner (p>0.05). A number of 36 features out of 46 were found to be less affected by scanner model.

Shape Features	Voxel Volume, Maximum3D Diameter, Mesh Volume, Major Axis Length, Shape, Sphericity, Least Axis Length, Elongation, Surface Volume Ratio, Maximum 2D Diameter Slice, Flatness, Surface Area, Minor Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row
First order	Median, Root Mean Squared, 90Percentile, 10Percentile, Mean
GLCM	Idmn, Idn, Correlation, Imc2, Imc1, MCC
GLDM	Dependence Non-Uniformity, Dependence Non Uniformity Normalized, Large Dependence High Gray Level Emphasis, Small Dependence Low Gray Level Emphasis.
GLRLM	Gray Level Non-Uniformity, Long Run Emphasis, Run Length Non Uniformity, Short Run Emphasis, Run Percentage, Run Length Non Uniformity Normalized
GLSZM	Small Area Low Gray Level Emphasis

Principle Component Analysis: PCA analysis results in 23 principle components accounting for all variation included in the entire dataset. However, the first 4 principle components revealed approximately 90.3% variation which is reasonably acceptable and can be potentially used to reduce curse of large data dimensionality. The scree plots of the all components are demonstrated in ***Figure (4) and Table (7)***.

It can be easily detected that the first 4 components bear large percentage of data variability. When plotting the sphere contribution versus the first two components, it appeared that PCA was able to discriminate between the two scanners significantly especially for largest and smallest spheres (11 and 12 versus 13 and 14, fig 5) that corresponds to sphere size of 1.1 mm and 1.0 mm versus 11.49 mm and 5.14 mm respectively.

Table (8) shows the first 5 contributing features of the top 4 principle components. It is well known that features of the same component are in linear correlation. Thus, it could be more useful to represent the whole dataset by choosing additional features from other components to increase the capability in accounting for more variation of the investigated dataset.

Table (7): Results of PCA. The eigen value of each component along with percent cumulative and variance percent. Notice that the first 4 components account for 90.3% of the overall variations.

Eigen Value	Variance	Percent cumulative	Variance Percent
PC.1	**58.1**	**61.8**	**61.8**
PC.2	**14.1**	**15.0**	**76.8**
PC.3	**9.4**	**10.0**	**86.8**
PC.4	**3.3**	**3.5**	**90.3**
PC.5	2.2	2.3	92.7
PC.6	1.8	1.9	94.5
PC.7	1.4	1.5	96.0
PC.8	1.1	1.2	97.2
PC.9	0.6	0.7	97.9
PC.10	0.6	0.6	98.5

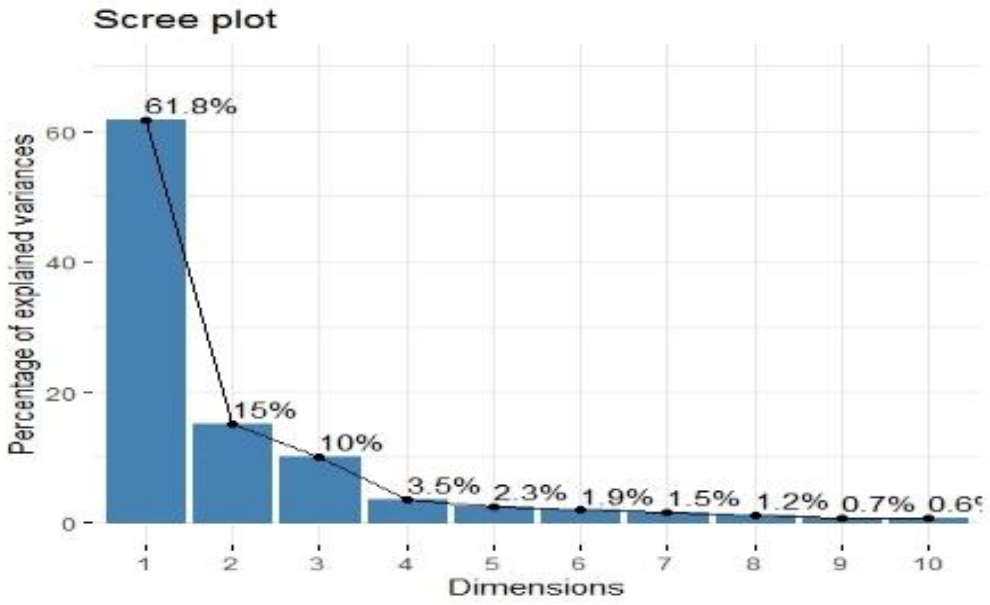


Figure (4): Principle component analysis revealed that the first three components can account for approximately 90.3 % of the variation of the entire dataset.

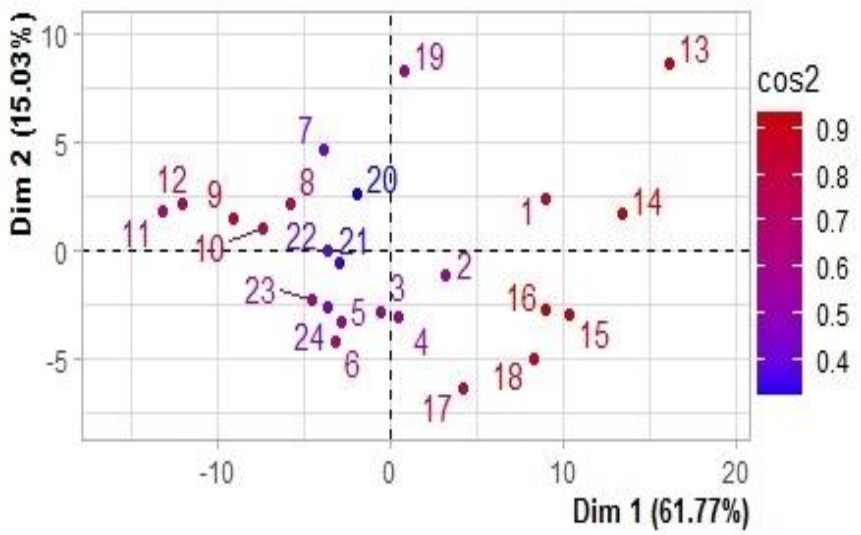


Figure (5): The contribution of the different spheres/lesions to the first two components. The 24 points refer to the 4 acquisitions performed by the two imaging systems at two different level of contrasts namely 4:1 and 10:1. The numbers 1-6 and 7-12 are for Siemens Biograph while the numbers 13-18 and 19-24 for high and low contrasts respectively. The PCA was able to distinguish between the spheres/lesions derived from the two scanners as shown by separating 13 and 14 apart from 11 and 12. The former two belongs to the Discovery STE while the latter two spheres belong to the Siemens Biograph.

Table (8): The top 4 features revealed from the first 4 principle components (PC). Since the features produced in each component are in correlation with their neighbors, they don't bear extra information in comparison to those features of other components and hence a single feature was selected from each component.

PC1	**Range**	**Maximum**	**SD**	**90 Percentile**	**MAD**
Feature Selected	Maximum (First order)				
PC2	Skewnes	**Idmn**	GLNU	GLNU	Idn
Feature Selected	Idmn (GLCM Matrix)				
PC3	**SZNUN**	Joint Entrop	MP	Coarseness	Busyness
Feature Selected	SZNUN (GLSZM Matrix)				
PC4	**SRLGL**	LGRLE	LGLE	LRLGE	Busyness
Feature Selected	SRLGLE (GLRLM Matrix)				
Abbreviations are described in Table (2).					

Clinical Correlations: The above analytical steps resulted in most appropriate features that can be used to maintain stability, reproducibility, and less deviation. However, the clinical correlations of texture features need to also to be identified and sorted out. A number of 11 patients with head and neck cancer were evaluated using all features outlined in ***Table (2)*** versus patient staging. An interesting finding was that there were only two features (out of 108 features) which showed correlations with patient TNM staging. These features were flatness and Maximal Correlation Coefficient (MCC) with Spearman correlation of 0.366 (p=0.044) and 0.380 (p=0.037) respectively. Moreover, the Kendall tau statistics provided close results specifically 0.448 (p=0.042) and 0.491 (p=0.024) respectively.

DISCUSSION:

Radiomics features have been of particular interest in the last decade providing useful imaging biomarkers in supporting clinical decision making. Several reports were published to report the diagnostic, prognostic as well as patient stratification for proper patient management [(16)]. However, the reproducibility, repeatability as well as stability of those feature was not widely investigated.

Image noise, limited spatial resolution and partial volume, scanner detector limitations, reduced count sensitivity and data acquisition parameters as well as different reconstruction methods are among the most crucial variables. Therefore, the process of extracting information either on human observer basis or physical/mathematical basis are confounded by one or more of those factors leading to unreliable outcome and inconclusive results [(19)].

The cancer imaging archive represent large but an increasing repository of pathological, anatomical and functional imaging dataset that may be accompanied with genomic or proteomic profiles of individual patients [(13)]. It comprises a wide array of standard phantoms for PET, CT and MR imaging modalities. PET was found to provide superior performance in comparison to CT radiomics models [(20)]. While PET has unique characteristics in providing valuable metabolic and physiologic information, it suffers from degrading factors that affect the final reconstructed images with variable impact on quality and diagnostic accuracy. However, radiomics provides a lot of information about the underlying tumor biology and thus can be used for tumor phenotyping and success tailored treatment [(20-22)].

From the 108-feature generated from the phantom scans, there were 56 features of variability and percent deviation less than 10% in both evaluation metrics. If high precision measurements are required in the range of 5%, then the number of features is reduced to only two features namely Inverse Difference Moment Normalized (Idmn) and Inverse Difference Normalized (Idn). However, when the 56 features were evaluated for the influence of sphere/lesion contrast the number was reduced to 46 features that are not affected by variation of low or high image contrast.

Moreover, there were 36 features out of 46, which showed no significant differences due to the use of different scanners in data acquisition. These three tests resulted in a reduction of total features from 100% to 33.3% which is highly desirable to avoid model over fitting.

Results of the PCA was so interesting and has made a significant selection to those feature that could account for large percentage of data variation, approximately 90.3% was obtained with the use of only 4 texture from different matrices. They were the SUV max, inverse difference moment normalized (Idmn), size zone non-uniformity normalized (SZNUN), and short run low gray level emphasis (SRLGLE). Those features were selected to represent the maximum variance of the data in question. The reason behind selecting the SUV max but not the "range" feature is due to the fact that SUV max is commonly used in routine practice of clinical oncology and its role has been identified in many malignancies. The reproducibility of the feature resulted in three features that are not largely affected by the sphere size, namely, Idn, Idmn and sphericity. This is an additional interesting finding since Idmn and Idn have already achieved less sensitivity in terms of CV, percent deviation, level of image contrast, scanner model besides sphere size. It is also consistent with previous reports [(23)].

In the present study we have used a semi-automated segmentation method as it was found to improve radiomics reproducibility in comparison to manual delineation [(24)]. In a previous report, it was found that first order statistics and shape features are more reproducible than texture features [(11)]. This is consistent with the finding presented here as most of features of high CV were those derived from texture matrices. Furthermore, first order statistics is less sensitive to image processing steps relative to texture features. However, the latter provide a more descriptive measure of voxels arrangement and their relative neighborhood to each other which is in line with the notion of tumor heterogeneity [(21)].

The opportunity to analyze the images provided by cancer imaging archive is unique since it allows researchers to harmonize their finding and reproduce the same results adding more evidence to the final conclusion. It also provides investigators to externally validate the performance of the proposed models and ensure stability. When correlating the 108 features with patient status, two features only showed significant correlation with tumor stage namely maximal correlation coefficient and flatness. Combining the two features with the other features derived from the other analytical and reproducibility results should provide a strong platform for future clinical studies aiming to derive a robust radiomics-based clinical model.

CONCLUSIONS:

There are several factors and variables that interfere to affect reproducibility and robustness of radiomics derived from PET data. A lot of efforts and attention need to be brought to tackle all those variables in order to exploit radiomics information to the maximum extent. The 108 features were reduced to a lower percentage of features while maintaining large percentage of variance and reproducibility. It has appeared that some but few features can serve as potential candidates for further radiomics analysis and clinical experimentation. While there were large number of features that can be used for clinical correlation, the present study has obtained two features that can correlate with tumor stage of patients with head and neck cancers. The combination of texture feature of robust technical qualifications to those of clinical value would finally improve the diagnostic or prognostic clinical model.

ACKNOWLEDGMENTS:

The author discloses that he has no conflict of interest.

REFERNCES:

1. ***Singh G.K and Jemal A,.*** Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States: Over Six Decades of Changing Patterns and Widening Inequalities. J. Environ Public Health; doi: 10.1155/2017/2819372. Epub; 2017.

2. ***O'Connor J.P, Aboagye E.O, Adams J.E, et al,.*** Imaging biomarker roadmap for cancer studies. Nat. Rev. Clin. Oncol. 14, 169-186; 2017.

3. ***Sossi V,.*** Advances in PET Methodology. Int. Rev. Neurobiol, 141, 3-30; 2018.

4. ***O'Connor J.P, Rose C.J, Waterton J.C, et al,.*** Imaging intratumor heterogeneity: role in therapy response, resistance and clinical outcome. Clin. Cancer Res. 21, 249-257; 2015.

5. ***Park J.E, Park S.Y, Kim H.J, et al.*** Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J. Radiol. 20, 1124-1137; 2019.

6. ***Zwanenburg A, Vallieres M, Abdalah M.A, et al.*** The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology, 295, 328-338; 2020.

7. ***Cortes-Rodicio J, Sanchez-Merino G, Garcia-Fidalgo M.A, et al,.*** Identification of low variability textural features for heterogeneity quantification of (18)F-FDG PET/CT imaging. Rev. Esp. Med. Nucl. Imagen. Mol. 35, 379-384; 2016.

8. ***Galavis P.E, Hollensen C, Jallow N, et al,.*** Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta. Oncol. 49, 1012-1016; 2010.

9. ***Rahmim A, Bak-Fredslund K.P, Ashrafinia S, et al,.*** Prognostic modeling for patients with colorectal liver metastases incorporating FDG PET radiomic features. Eur. J. Radiol. 113, 101-109; 2019.

10. ***Manafi-Farid R, Karamzade-Ziarati N, Vali R, et al,.*** 2-[(18)F]FDG PET/CT radiomics in lung cancer: An overview of the technical aspect and its emerging role in management of the disease Methods, Jun. 1;S1046-2023(19)30325-1; 2020.

11. ***Traverso A, Wee L, Dekker A, et al.*** Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int. J. Radiat. Oncol. Biol. Phys. 102, 1143-1158; 2018.

12. https://imaging.cancer.gov. Accessed on 1st August; 2020.

13. https://www.cancerimagingarchive.net Accessed on 1st, August; 2020

14. ***Liu X, Comtat C, Michel C, et al,.*** Comparison of 3-D reconstruction with 3D-OSEM and with FORE+OSEM for PET. IEEE. Trans. Med. Imaging, 20, 804-814; 2001.

15. ***Beichel R.R, Van Tol M, Ulrich E.J, et al,.*** Semiautomated segmentation of head and neck cancers in 18F-FDG PET scans: A just-enough-interaction approach, Med. Phys. 43, 2948-2964; 2016.

16. ***Chen B, Yang L, Zhang R, et al,.*** Radiomics: an overview in lung cancer management-a narrative review. Ann. Transl. Med. 8, 1191; 2020.

17. https://www.slicer.org/. Accessed on 1st August; 2020.

18. https://cran.r-project.org/. Accessed on 1st August; 2020.

19. ***Nardone V, Reginelli A, Guida C, et al,.*** Delta-radiomics increases multicentre reproducibility: a phantom study. Med. Oncol. Mar. 31;37 (5):38; 2020.

20. ***Bogowicz M, Riesterer O, Stark L.S, et al,.*** Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncol, 56, 1531-1536; 2017.

21. ***Aerts H.J, Velazquez E.R, Leijenaar R.T, et al,.*** Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun, Jun. 3;5:4006; 2014.

22. ***Arshad M.A, Thornton A, Lu H, et al,.*** Discovery of pre-therapy 2-deoxy-2-(18)F-fluoro-D-glucose positron emission tomography-based radiomics classifiers of survival outcome in non-small-cell lung cancer patients. Eur. J. Nucl. Med. Mol. Imaging, 46, 455-466; 2019.

23. ***Altazi B.A, Zhang G.G, Fernandez D.C, et al,.*** Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J. Appl. Clin. Med. Phys. 18, 32-48; 2017.

24. ***Parmar C, Rios Velazquez E, Leijenaar R, et al,.*** Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS. One, Jul 15;9 (7):e102107; 2014.