Quantification of SPECT concentric ring artifacts by radiomics and radial features

Emilio Mezzenga
IRST: Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori Srl

Anna Samelli
IRST: Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori Srl

Giovanni Bellomo
IRST: Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori Srl

Frank P DiFilippo
Cleveland Clinic

Christopher J Palestro
Donald and Barbara Zucker School of Medicine at Hofstra/Northwell

Kenneth J Nichols (knichols@northwell.edu)
Northwell Health https://orcid.org/0000-0003-2010-7078

Original research

Keywords: Image artifacts, SPECT, gamma camera, quality assurance, image texture analysis, inter-observer agreement, phantom

DOI: https://doi.org/10.21203/rs.3.rs-78752/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

**Background:** Conspicuous concentric ring artifacts in phantom reconstructions triggers retuning SPECT systems. These evaluations are visual, not quantitative. Our study was undertaken to determine the degree to which observers agree about SPECT concentric ring artifacts, and to test whether quantitative texture analysis metrics correspond to significant artifacts.

**Methods:** Test data were acquired as part of quarterly quality assurance using standardized SPECT phantoms containing solid spheres, solid rods and volumes of uniform activity concentration loaded with $^{99m}$Tc. Forty SPECT studies were identified as having concentric ring artifacts or were acquired to assess whether artifacts were resolved following camera retuning after obtaining an unacceptably non-uniform result. Transaxial reconstructions were reviewed independently by two medical physicists who graded severity of artifacts on a 5-point scale. Counts were tabulated in volumes of interest created in uniform phantom sections, from which were computed 72 radiomics image texture analysis metrics. Radial contrast ($R_{\text{Contrast}}$) derived from the radial profile of summed slices transformed into polar coordinates and radial noise-to-signal ($R_{\text{NSR}}$) also calculated.

**Results:** Artifacts were considered sufficiently severe to require camera retuning in 10 rods sections, 17 sphere sections, and 16 uniform sections. In uniform sections, there was “good agreement” for inter-observer and intra-rater assessments ($\kappa = 0.66$, Fisher exact $p < 0.0001$ and $\kappa = 0.61$, Fisher exact $p = 0.001$, respectively). While 3 radiomics image analysis features agreed significantly ($p = 0.001$) with visual detection of significant artifacts in uniform sections, the parameters most strongly associated with severe artifacts were $R_{\text{Contrast}} > 4.75\%$ and $R_{\text{NSR}} > 2.7\%$, for which ROC AUC accuracy = $88\%\pm5\%$, sensitivity = $83\%$, specificity = $83\%$, $p < 0.0001$. Accuracy was $76\%-78\%$ for the 3 radiomics features, with significantly lower specificity ($48\%-61\%$, $p < 0.05$) than $R_{\text{Contrast}}$ and $R_{\text{NSR}}$. Increasing magnitude of $R_{\text{Contrast}}$ and $R_{\text{NSR}}$ correlated significantly with increasingly severe artifact scores ($\rho = 0.71-0.72$, $p < 0.0001$).

**Conclusion:** There was good agreement among physicists as to the presence of circular ring artifacts in uniform sections of SPECT quality assurance scans, with artifacts accurately detected by radial contrast and noise-to-signal ratio measurements.

**Background**

Numerous factors influence the accuracy of single photon emission computed tomography (SPECT) reconstructions. Besides patient motion issues (e.g., tidal breathing motion and cardiac contraction), complications include photon attenuation and scatter, variability of point spread function with distance from detectors, imperfect detector efficiency, and low counting statistics [1]. Over the years considerable efforts have been applied to address these issues through the application of physics to optimize the reconstruction process and electric and mechanical engineering to improve SPECT system designs. Yet, data acquisition problems can occur abruptly, or as a result of gradual electrical or mechanical
deterioration of SPECT system components, which can compromise the reliability of acquired data for clinical use [2].

For this reason, adherence to regularly scheduled standardized quality assurance (QA) procedures is important in order to identify problems with data acquisitions before they can potentially result in inaccurate clinical readings. These QA procedures include quarterly SPECT phantom reconstructions, which typically are assessed visually for contrast, spatial resolution, and uniformity of response [3].

In the course of evaluating SPECT phantom transaxial reconstructions to assess integrity of acquired image data, the appearance of conspicuous concentric ring artifacts often will trigger retuning conventional rotating SPECT detectors, as those artifacts may be due to inadequate uniformity corrections [4]. One cause of this can be insufficient count density of the correcting flood field, the details of which have been studied extensively [5]. It has been shown that choice of tomographic reconstruction filters affects the appearance of artifacts, and that perceptibility of artifacts depends on their size and position within the object [6].

However, these evaluations are visual, not quantitative, and criteria have not been established for deciding what constitutes an artifact that is sufficiently severe to necessitate detector retuning. When medical imaging physicists convene to read QA studies from other institutions, they begin with training data sets to calibrate their readings and to minimize perceptual differences among them, and they read by consensus. However, different viewing environments, such as brightness, contrast and screen resolution of display monitors, can influence visual impressions. Also, technologists who acquire QA data may not be as well versed in these types of artifacts may be uncertain as to just how serious a potential artifact could be.

To address these issues, algorithms have been devised to analyze standardized SPECT QA phantoms, both to report the results of spatial resolution and contrast results [7,8], and to report the possibility of artifacts [9]. Some of these algorithms are based on texture analysis [7,9]. Approaches to quantifying the severity of circular ring artifacts have included annular sampling techniques [10], or by the use of the Hough transform, the efficacy of which to detect artifacts has been assessed using computerized Monte Carlo phantom simulations [11].

Our study was conducted with two goals: (1) to determine how well medical physicists who are experienced in making these evaluations agree with one another about the presence and severity of concentric ring artifacts, and (2) to establish if there are any quantitative image metrics that correspond to visual judgments as to the significance of these artifacts. Having a quantitative parameter could help less experienced observers decide whether to escalate a remediation process when confronted with an imperfect-appearing phantom reconstruction result.

**Methods**
**Phantom data acquisitions**

All test data were acquired at one institution as part of routine quarterly QA using standardized SPECT phantoms. The standardized phantom included a cylindrical water bath, with Plexiglas® inserts of 6 rod sizes, solid spheres of 6 sizes, and a uniform volume devoid of rods or spheres $^{[12,13]}$. Phantoms were loaded with 666 MBq-740 MBq $^{99m}$Tc. Each 360° set of projection data were comprised of 128 projections acquired as 128x128 matrices for 32-36 x $10^6$ counts. Magnification factors of 1.00 to 1.85 were used (pixel size = 3.3±0.7 mm; range = 2.1 to 5.2 mm).

All data sets were reconstructed by filtered backprojection with a Hanning post-filter (cutoff = 1.0 cm$^{-1}$), and corrected by a simple Chang attenuation correction with attenuation coefficient of $\mu = 0.11$ cm$^{-1}$, or as adjusted as necessary in order to obtain the most uniform appearance of transaxial reconstructions of the volume of the phantom containing no rods or spheres.

From a collection of over 200 acquisitions performed for 12 different SPECT systems between 2016-2020, 40 studies were identified by a medical physicist at the time of data acquisition as either having evidence of concentric ring artifacts, or else were acquired to determine whether concentric ring artifacts were resolved following camera retuning performed within 1 week after obtaining an unacceptably non-uniform-appearing result. A 41$^{st}$ phantom study also was considered as a reference standard for which observers agreed that no artifacts were visible. At the time of these QA phantom data acquisitions, camera ages ranged from 20 years old to 1 month old. Ten systems were dual-detector systems and 2 were single-detector systems.

**Visual phantom readings**

Automated algorithms written in IDL 8.4 were run on all transaxial phantom data that generated a series of output jpg image files $^[7]$. Because some accrediting agencies request displays of all transaxial slices $^[3]$, multiple jpg files were generated per phantom for all transaxial slices spanning the entire height of the phantom, and a separate composite jpg file was generated showing a composite of 3 images (Fig. 1): a summed rod section, the single section through spheres with the highest contrast, and the section of uniform activity concentration. The jpg files were reviewed and scored by observers as to the presence and severity of concentric ring artifacts.

To gauge inter-rater agreement of the presence and severity of concentric artifact, two experienced medical physicists (KJN and FPD) independently scored severity of artifacts on a 5-point scale (0 = “no artifact”; 1 = “probably insignificant artifact”; 2 = “equivocal”; 3 = “probably significant artifact”; 4 = “severe artifact definitely requiring camera retuning”), separately for rod sections (Fig. 2), sphere sections (Fig. 3) and uniform phantom sections (Fig. 4). Single, separate, scores were generated by both readers for the perception of artifacts as seen on the composite summary images (Fig. 1).

One physicist (KJN) re-scored the images blinded to his initial scores, to enable assessment of intra-rater reproducibility. Readers had no knowledge of any texture analysis metrics values, as these were not
computed until after all readings were completed for both physicists.

Following the assignment of all scores to all phantoms by both physicists, readings also were
dichotomized so that any reading > 2 was assigned 1 and were otherwise 0.

**Radiomic features extraction**

Texture analysis was performed by two medical physicists (EM and AS) on DICOM files of the transaxial
reconstructions that were transmitted electronically to them in Italy. SPECT datasets were imported into
MIMvista software (MIM Software, Inc. Cleveland, OH). Using the open source CGITA software \([14]\), they
manually positioned cylindrical volumes of interest (VOI) to sample counts of the transaxial phantom
sections of uniform activity concentration, in the region of the phantom containing no rods or spheres
(Fig 1) When visible, VOIs were centered on a perceived center of a concentric ring artifact, individually for
each phantom data set; otherwise, they were centered on the middle of a transaxial section. The CGITA
algorithms enabled the computation of 72 image texture features (Table 1), which included image texture
metrics based on Gray Level Co-occurrence Matrix (GLCM) computations \([15,16]\), including image energy
and entropy measures that have been useful in tumor characterization studies \([17]\). A GLCM is
a matrix that defines the distribution in an image of co-occurring count levels per voxel at a given voxel
offset. In addition to, but distinct from, GLCM metrics, texture feature coding metrics were also computed
by the CGITA algorithms (Table 1) \([18]\), which have been used in ultrasonography to categorize liver
disease \([19]\).

Among other additional metrics, the CGITA algorithms computed 11 different types of features of voxel-
alignment metrics (Table 2) and 11 intensity size zone metrics (Table 3) \([9]\). Voxel-alignment features are
useful in tumor characterization studies \([20]\), and are based on the construction of a Gray Level Run
Length Matrix (GLRLM) for statistical image texture characterization that consists of counting the
number of voxels having the same count level in a given direction \([21]\). CGITA analysis parameters were
selected to establish the size of GLCM voxel neighborhoods and the extent and orientation angles of the
GLRLMs as follows: (1) local normalization was performed with minimum and maximum count values
calculated inside the VOI; (2) the distance between pairs of voxels was set to 1 voxel (thereby specifying
that computations were to be performed for directly neighboring voxels); (3) the angular orientation was
averaged; and (4) full range of scaled count values was set to 64 bins.

**Radial features extraction**

Taking into account the radial nature of the concentric ring artifacts, two more image features were
computed: radial contrast \((R_{\text{Contrast}})\) and radial noise-to-signal ratio \((R_{\text{NSR}})\) \([22]\). \(R_{\text{Contrast}}\) and \(R_{\text{NSR}}\) were
computed from the 41 SPECT phantom images \(f(x,y,z)\) using the “ImageJ” software \([23]\). The workflow
adopted to obtain radial features is illustrated in Fig. 5. Depending on voxel dimension, 8-10 tomographic
homogeneous phantom transaxial sections \((z_{th})\) were summed together:
Summed images $f_z(x,y)$ were transformed to polar coordinates $f_z(\rho,\phi)$, where:

$$\rho = \sqrt{x^2 + y^2}$$

and:

$$\phi = \tan^{-1}\left(\frac{y}{x}\right)$$

To obtain the radial profile $f_z(\phi)(\rho)$, the images $f_z(\rho,\phi)$ were summed over the $\phi$ angular variable:

$$f_z(\rho) = \sum_{\phi} f_z(\rho,\phi)$$

This profile was used to compute $R_{\text{Contrast}}$ and $R_{\text{NSR}}$:

$$R_{\text{Contrast}} = \frac{\max(f_z(\phi)(\rho)) - \min(f_z(\phi)(\rho))}{\max(f_z(\phi)(\rho)) + \min(f_z(\phi)(\rho))}$$

$$R_{\text{NSR}} = \frac{\sigma(f_z(\phi)(\rho))}{\mu(f_z(\phi)(\rho))};$$

with $\mu$ and $\sigma$ being the mean value and the standard deviation of the radial profile $f_z(\phi)(\rho)$, respectively.

**Features analysis**

In addition to metrics being computed individually for each phantom data set, texture features were also compared to a reference standard image data set. To do so, all metrics were computed for a reference phantom that had been identified as being *a priori* uniform, i.e., graded as “0” independently by both physicists for all rod, sphere, and uniform sections. Relative differences then were computed between the reference standard (i.e. the artifact-free phantom data set) and each of the 40 test data sets.

**Statistical analysis**

Statistical analysis was performed using MedCalc software [24]. Values are reported as means ± one standard deviation. Whether or not continuous variables were normally distributed was determined by the Shapiro-Wilk test. Trends of variables with increasing severity of concentric ring artifacts were quantified by $\chi^2$ rank correlation with Spearman's $\rho$. Significance of mean differences was assessed by ANOVA, which included computation of F-ratios (measuring variations among groups for which a value of 1.0
results for the null-hypothesis). The most discriminant of the 72 radiomic features were identified by means of the logistic least absolute shrinkage and selection operator (LASSO) approach [25]. LASSO is a regression analysis method that both selects discriminant variables and performs regularization in order to enhance the accuracy of model predictions and interpretability of the statistical models that are produced. The LASSO operator for our application used the dichotomous classification of the 40 phantom images, and a shrinkage penalty parameter (\( \lambda \)) [26]. Inter-rater agreement and intra-rater reproducibility of visual scores was determined by the kappa statistic [27], with significance determined by Fisher’s exact test. Dichotomized visual scores were formed as > 2 of the mean of the 2 readers’ severity grades of the uniform phantom sections. Logistic regression determined which of the radiomic and radial features were able to discriminate positive from negative dichotomous visual scores.

Receiver operating characteristics (ROC) analysis determined the accuracy of each parameter to agree with dichotomous visual scores, producing the threshold for discrimination (Youden index), sensitivity and specificity for each parameter. Comparison of proportions was performed by the \( \chi^2 \) test. Significance of differences among ROC AUC results was assessed [28]. A probability (p) < 0.05 was considered to be statistically significant.

Results

Visual scores

Count levels were similar among phantom acquisitions, with \((34\pm1) \times 10^6\) counts, and were not normally distributed (kurtosis = 6.0, p = 0.0004), as they were narrowly clustered near \(34 \times 10^6\) counts, so that differences among phantoms of counts or of noise-to-signal ratios were unlikely to account for perceived artifacts.

Among the 40 phantoms, artifacts were considered sufficiently severe to require camera retuning based on the visual scores for 10 rods sections, 16 sphere sections, and 21 uniform sections for the first reader and for 10 rods sections, 17 sphere sections, and 16 uniform sections for the second reader. Thus, the percent of phantoms in which artifacts were perceived were smaller in rod than in uniform sections (24% versus 39%-51%, p = 0.01). From the displays of all individual transaxial sections, inter-observer comparisons demonstrated “fair agreement” as to artifacts for rods (\( \kappa = 0.33, p = 0.09 \)), “good agreement” for spheres (\( \kappa = 0.64, p < 0.0001 \)), and “good agreement” for uniform sections (\( \kappa = 0.65, p < 0.0001 \)). Correlation between readers was stronger for sphere sections and uniform sections than for rod sections (p = 0.004).

Intra-observer reproducibility demonstrated “moderate agreement” for rods (\( \kappa = 0.58, p = 0.0001 \)), “good agreement” for spheres (\( \kappa = 0.70, p < 0.0001 \)), and “good agreement” for uniform sections (\( \kappa = 0.70, p < 0.0001 \)).
For the perception of artifacts seen on composite images, there was “moderate agreement” for interobserver comparisons (κ = 0.47, Fisher exact p = 0.01), and “good agreement” for intra-observer reproducibility (κ = 0.68, p < 0.0001).

**Correlation with increasing artifact severity**

The mean of both observers’ scores of uniform phantom sections were used to establish 5-level consensus readings of artifact severity, rounded up for non-integer mean scores. Of all image metrics tested, $R_{\text{Contrast}}$ had the highest Spearman coefficient with increasing magnitude of $R_{\text{Contrast}}$ correlated to increasing severity of artifacts ($\rho = 0.72, p < 0.0001$) (Fig. 6). $R_{\text{NSR}}$ also had significant correlation with increasing artifact severity ($\rho = 0.71, p < 0.0001$).

**Image analysis metrics**

Both radiomic and radial features were normally distributed (p-value > 0.05). For all of them the absolute value of the difference between the reference standard and the 40 test data sets was in stronger agreement with dichotomous visual scores than the original metrics themselves. Consequently, the results reported in Tables 4–6 are for absolute values of differences from the reference standard phantom.

Among all tested parameters, $R_{\text{NSR}} > 2.7\%$ and $R_{\text{Contrast}} > 4.75\%$ had the highest accuracy (ROC AUC = 88±5%, p < 0.0001), and similar sensitivity and specificity of 83% (Table 4), with dichotomous accuracy of 83% for discriminating between positive and negative mean dichotomous scores. In addition, LASSO indicated that 3 radiomic features (i.e. 1 Intensity-size-zone feature and 2 voxel-alignment features) significantly predicted dichotomous visual artifact scores, with a shrinkage penalty parameter $\lambda = 0.11$, corresponding to the cross-validation error between the model and the dichotomized visual scores [25]. For these 3 radiomic features, ROC analysis indicated $p < 0.05$ (Fig. 7). For a ROC result being significantly different from chance (ROC AUC = 50%), the metric’s AUC would have to have been > 70%. While ROC AUC and sensitivity values were not statistically different among the parameters reported in Table 4, specificity was significantly lower for the 3 radiomic features than for $R_{\text{NSR}}$ and $R_{\text{Contrast}}$.

Moreover, $R_{\text{NSR}}$ and $R_{\text{Contrast}}$ had the highest ROC AUC values (Table 4).

The metric with the strongest association (highest logistic regression $\chi^2$ value) with visual scores of artifacts seen in uniform sections was $R_{\text{NSR}}$, for which the logistic regression model correctly predicted 83% of cases, with similar results for $R_{\text{Contrast}}$ (Table 5).

$R_{\text{NSR}}$ and $R_{\text{Contrast}}$ had significantly higher values for phantoms with severe uniformity concentric ring artifacts than those without artifacts (Table 6, Fig. 8). Consistent with ROC and logistic regression analyses results, the F-ratios were higher, and the probability values lower, for $R_{\text{NSR}}$ and $R_{\text{Contrast}}$ than for the radiomic features (Table 6).
Discussion

As with any type of device used to collect data that contributes to forming a clinical diagnosis, it is important to assess the performance characteristics of rotating SPECT systems. SPECT equipment testing includes the acquisition and analysis of three-dimensional phantoms to gauge system tomographic spatial resolution, image contrast, and uniformity of response, and it is recommended to be performed as part of acceptance testing \[29\], and thereafter on a quarterly basis \[3\]. Failure to recognize concentric ring artifacts can contribute to misinterpretation of nuclear cardiology studies \[30\], and non-cardiac nuclear medicine studies \[5\].

Establishing whether a particular image feature is visible is non-trivial. One of the goals of our investigation was to document the extent to which two independent medical imaging physicists agree on the appearance and severity of artifacts. Thresholds of visible detection in a $34 \times 10^6$ count phantom study and in a clinical study might be different. Experience of the observer is a factor, and evaluations even by experienced physicists might differ. Visibility indexes can help in this regard, such as those dependent on feature contrast and feature dimensions \[31\], but ultimately such indices require calibration with scores assigned by human observers \[32\].

We found that readers agreed more successfully when observing all of the transaxial sections than when viewing just the summary images, and that our readers had “good agreement” in detecting artifacts in uniform and sphere sections but only “fair agreement” about rods sections. $\kappa$ values of reader agreement and reading reproducibility were higher for uniform sections and for sphere sections than for rod sections. That may be due to the fact that there is less distraction in uniform sections in perceiving ring artifact patterns compared to rod sectors, for which such patterns are superimposed on count variations within and between rod images of different sizes.

Visibility of a concentric ring artifact is one thing, but deciding that an artifact is sufficiently severe to cause potential clinical reading problems is another matter, because that requires removing a camera from clinical service long enough to do a thorough re-tuning of the detectors. The process of camera re-tuning may be a fairly unobtrusive process, such as routine detector tuning or flood calibration by a technologist of $100-120 \times 10^6$ counts, or can be more time-consuming, involving a full calibration of energy, linearity, and uniformity corrections by a field service engineer. Consequently, having a quantitative basis for triggering the decision to retune detectors would be useful information for technologists performing quarterly SPECT quality assurance scans.

While the results of quality assurance phantom scans usually are assessed visually, recent progress has been reported in quantifying scanner performance automatically \[7,8,9,11\]. Among the parameters that are evaluated in SPECT phantom scans, non-uniformity is perhaps the most challenging. O’Connor et al \[6\] found that choice of reconstruction filters has a pronounced effect on the manifestation of planar flood field non-uniformity on the formation of tomographic concentric ring artifacts, and that the ability of observers to perceive such artifacts depends on the size and location of these artifacts; smaller artifacts
closer to the center-of-rotation were less perceptible than larger ones further from the center [6]. That is consistent with a reader tending to give more credence to a pattern perceived to be repeated over more pixels than over fewer pixels, as there would be a tendency to give more weight given to a similar pattern confirmed in multiple locations, even when these deviations from neighboring counts are small. Concentric ring artifacts constitute a type of pattern, and as such, the bigger the radius, the more likely a human observer will detect it in the first place, even if subtle. The larger the size of the pattern, the more likely a person is to perceive it, as the suggestion of a pattern in one area of a field can then be reinforced in neighboring areas, and the more areas that confirm the same impression, and that conform to a continuation of a pattern in neighboring areas, the higher the confidence of the reader that the pattern is genuine.

An advantage of the polar coordinate transformations used in our study, and in the application of the Hough transform implemented by Hirtl et al [11], is the fact that non-uniformities tend to be concentric rings, so that ring artifacts of any size become mapped into straight lines across the entire extent of the image matrix in polar coordinates. Polar coordinate transformations have the effect of bringing concentric artifacts that are near to the center-of-rotation up to an equal footing with those at a greater radius. This process relies on knowing the center of the reconstruction matrix (i.e. if an offset was applied during image reconstruction). A possible area of further study could be the implementation of a search algorithm that maximizes radial contrast ($R_{\text{Contrast}}$) and/or radial noise-to-signal ($R_{\text{NSR}}$), in an attempt to locate this matrix center.

Using ROC analysis, we independently found that a radial contrast value > 4.75% confirmed findings of other investigations that > 4% of the amplitude of Hough-transformed transaxial images above background counts agreed optimally with medical physicists as to the seriousness of concentric ring artifacts [11]. The fact that the $R_{\text{Contrast}}$ and $R_{\text{NSR}}$ parameters were more effective at agreeing with our medical physicists scores than any of the 72 radiomics features likely can be explained by the fact that the radial parameters make use of the \textit{a priori} condition that these artifacts are circular patterns, whereas the radiomics features are more generic and shape-independent.

**Limitations**

To maximize spatial resolution, patient data usually are acquired with contouring, not as circular orbits. Our acquisitions were performed with contouring, and while some acquisitions may have been of a fixed radius, they were not necessarily so. Considering that the phantom we used was a simple cylinder, these orbits likely were close to being circular. It has been shown that non-uniform planar flood field corrections create circular artifacts in SPECT transaxial reconstructions for orbits of a fixed radius, but more complicated shapes when elliptical orbits are used [5].

Our study was limited to documenting the degree to which medical physicists agree with one another in their perception of the presence and severity of artifacts, not in whether there were problems in actual clinical studies. By linking $R_{\text{NSR}}$ and $R_{\text{Contrast}}$ to their visual impressions, we have provided a means to
connect quantified imaging parameters computed from standardized phantom scans to whether or not an artifact would have been judged to be sufficiently serious to require service on a SPECT system. This does not guarantee that a SPECT system for which \( R_{Contrast} > 4.75\% \) would invariably result in erroneous diagnoses in all clinical studies subsequently acquired with this equipment. There are innumerable variables involved in forming a final clinical diagnosis. Our results just serve as a guideline for acting on the results of phantom scans.

Our analyses of radiomic features and radial parameters were performed only in uniform sections, not in sphere or rod transaxial sections. The fact that inter-rater and intra-rater agreement were lowest for rod sections suggest that it is more difficult to decide about the presence and severity of artifacts in those sections compared to sphere and uniform phantom sections. Yet, we found the concentric ring artifacts were detected by both readers throughout the individual phantom sections. The fact that some radiomic features agree well with visual impressions in uniform sections suggests that further work is warranted to enable texture analysis applications to non-uniform phantom areas containing spheres and rods, where artifacts also appear.

It is possible that re-sampled images or images acquired with different spatial resolutions may have produced different ROC results from those reported in Table 4 for F1, F2 and F3. Because of the linear nature by which the radial measures were obtained in our investigation, it is plausible that \( R_{NSR} \) and \( R_{Contrast} \) should be relatively less sensitive to resampling issues and relatively less affected by different acquisition parameter choices than the other radiomics features we computed. The values of textural radiomic features can be affected by acquisition and reconstruction parameters, as reported for CT and PET imaging \([^{33,34,35}]\), and these influences require further investigation for SPECT imaging.

**Conclusion**

There is good agreement among observers, with reproducible results, as to the presence of circular ring artifacts in uniform sections of SPECT quality assurance scans, and some texture analysis metrics agree well with visual impressions in uniform sections. From among all computed image metrics, radial contrast, derived from the radial profile of summed slices transformed into polar coordinates, and radial noise-to-signal measurements demonstrated the highest accuracy of 88% to identify severe concentric ring artifacts in uniform sections of test phantom data. These quantified variables thus have the potential to replace subjective visual evaluation, thus promoting the automation of SPECT quality assurance and aiding the assessment of the degree to which efforts are successful in remedying detected equipment deficiencies.

**Abbreviations**

AUC area under curve

GLRLM Gray Level Run Length Matrix
Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: The data sets used for the current study are available for the corresponding author on reasonable request.

Competing interests: None of the authors have any competing interests to report relevant to this investigation.

Funding: Not applicable.

Author’s contributions: EM, AS, GB and KJN created and executed image analysis algorithms and performed statistical analysis of results. FPD and KJN performed visual readings. All authors contributed meaningfully to study design and to writing the manuscript. All authors have read and approved the final manuscript.

Acknowledgements: The authors Fritzgerald Leveque, May Liu, and Christopher Caravella for their invaluable assistance in collecting and processing the phantom data analyzed in this investigation.

References

1. Rosenthal S, Cullom J, Hawkins W, Moore SC, Tsui BMW, Yester M. Quantitative SPECT imaging: a review and recommendations by the Focus Committee of the Society of Nuclear Medicine Computer and Instrumentation Council. J Nucl Med. 1995;36:1489-513.

2. Celler A, Dixon KL, Chang Z, Blinder S, Powe J, Harrop R. Problems created in attenuation-corrected SPECT images by artifacts in attenuation maps: a simulation study. J Nucl Med. 2005;46:335-43.
3. https://ACR.org/Clinical-Resources/Accreditation. Accessed February 2, 2021

4. Leong LK, Kruger RL, O’connor MK. A comparison of the uniformity requirements for SPECT image reconstruction using FBP and OSEM techniques. J Nucl Med Technology. 2001;9;79-83.

5. Gullberg GT. An analytic approach to quantify uniformity artifacts for circular and noncircular detector motion in single photon emission computed tomography imaging. Med Phys. 1987;14:105-14.

6. O’Connor MK, Vermeersch C. Critical examination of the uniformity requirements for single-photon emission computed tomography. Med Phys. 1991;18:190-7.

7. Nichols KJ, DiFilippo FP, Palestro CJ. Texture analysis for automated evaluation of Jaszczak phantom SPECT system tests. Med Phys. 2019;46(1):262-72.

8. DiFilippo FP. Assessment of PET and SPECT phantom image quality through automated binary classification of cold rod arrays. Med Phys. 2019; 46:3451-61.

9. Samelli A, Mezzenga E, Vagheggini A, Piccinini F, Feliciani G, Belli ML, et al. Texture analysis in 177Lu SPECT phantom images: Statistical assessment of uniformity requirements using texture features. PLoS ONE. 2019;14(7):e0218814.

10. Madsen MT. A method for quantifying SPECT uniformity. Med Phys. 1997;24:1696-700.

11. Hirtl A, Bergmann H, Knausl B, Beyer T, Figl M, Hummel J. Technical Note: Fully-automated analysis of Jaszczak phantom measurements as part of routine SPECT quality control. Med Phys. 2017;44:1638-45.

12. Jaszczak RJ. United States Patent No. 4,499,375; 1985.

13. http://www.spect.com/pub/Flangeless_Jasczack_Phantoms.pdf. Accessed July 27, 2018.

14. Fang YH, Lin CY, Shih MJ, Wang HM, Ho TY, Liao CT, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. Biomed Res Int. 2014;2014:248505. doi:10.1155/2014/248505

15. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;SMC-3:610–21.

16. Cunliffe AR, Al-Hallaq HA, Labby ZE, Pelizzari CA, Straus C, Sensakovic WF, et al. Lung texture in serial thoracic CT scans: Assessment of change introduced by image registration. Med Phys. 2012;39:4679-90.

17. Dong X, Xing L, Wu P, Fu Z, Wan H, Li D, et al. Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-
fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumor stage. Nuc Med Commun. 2012;34:40-6.

18. Horng M-H, Sun Y-N, Lin X-Z. Texture feature coding method for classification of liver sonography. Comput Med Imaging Graph. 2002;26:33-42.

19. Gaber A, Hamdy A, Abdelaal HM, Elkattan A, Elshourbagy MM, Youness HA. Automatic classification algorithm for diffused liver diseases based on ultrasound images. IEEE Access. 2021; doi. 10.1109/ACCESS.2021.3049341; accessed on February 2, 2021.

20. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run length for texture analysis. Pattern Recogn Lett. 1990;11:415-9.

21. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, et al. Texture indexes and gray level size zone matrix. Application to cell nuclei classification. 10th International Conference on Pattern Recognition and Information Processing, PRIP 2009, 2009, Minsk, Belarus. pp.140-5. ⟨hal-01499715⟩

22. Michelson, AA. Studies in Optics. (University Press, 1927).

23. Rasband WS. ImageJ. U. S. National Institutes of Health, Bethesda, Maryland, USA, https://imagej.nih.gov/ij/, 1997-2018.

24. Medcalc Statistical Software version 19.3.1 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2020).

25. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Statist Soc B. 1996;58:267–88.

26. Friedman JM, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descendent. J Stat Softw. 2010:33;1-22.

27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45.

29. Busemann Sokole E, Plachcinska A, Britten A. Acceptance testing for nuclear medicine instrumentation. Eur J Nucl Med Mol Imaging. 2010;37:672-81.

30. DePuey EG. How to Detect and Avoid Myocardial Perfusion SPECT Artifacts. J Nucl Med. 1994;35:699-702.

31. Rose A. The sensitivity performance of the human eye on an absolute scale. J Opt Soc Am. 1948;38:196–208.
32. Burgess AE. The Rose model, revisited. J Opt Soc Am A. 1999;16:633-46.

33. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2(1):1-8.

34. Shaq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44(3):1050-62.

35. Shaq-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep. 2018;8(1):1-9.

Tables

**Table 1** CGITA software radiomics classes and features [15].

| Number of computed features | Radiomic class             | Feature measure                                                                 |
|-----------------------------|-----------------------------|--------------------------------------------------------------------------------|
| 7                           | Co-occurrence               | second angular moment, contrast, entropy, homogeneity, dissimilarity, inverse difference moment, correlation |
| 11                          | Voxel Alignment             | short-run emphasis, long-run emphasis, intensity variability, run-length variability, run percentage, low-intensity run emphasis, high-intensity run emphasis, low-intensity short-run emphasis, high-intensity short-run emphasis, low-intensity long-run emphasis, high-intensity long-run emphasis |
| 5                           | Neighborhood Intensity Difference | coarseness, contrast, busyness, complexity, strength |
| 11                          | Intensity Size Zone         | short-zone emphasis, large-zone emphasis, intensity variability, size-zone variability, zone percentage, low-intensity zone emphasis, high-intensity zone emphasis, low-intensity short-zone emphasis, high-intensity short-zone emphasis, low-intensity large-zone emphasis, high-intensity large-zone emphasis |
| 6                           | Normalized Co-occurrence features | second angular moment, contrast, entropy, homogeneity, dissimilarity, inverse difference moment, correlation |
| 13                          | SUV Statistics features     | minimum SUV, maximum SUV, mean SUV, SUV variance, SUV SD, SUV skewness, SUV kurtosis, SUV skewness (bias corrected), SUV kurtosis (bias corrected), TLG, tumor volume, entropy, SUV peak |
| 2                           | Texture Spectrum            | max spectrum, black-white symmetry |
| 4                           | Texture Feature Coding      | coarseness, homogeneity, mean convergence, variance |
| 8                           | Texture Feature Coding Co-occurrence | second angular moment, contrast, entropy, homogeneity, intensity, inverse difference moment, code entropy, code similarity |
| 5                           | Neighboring Gray-Level Dependence | small-number emphasis, large-number emphasis, number non-uniformity, second moment, entropy |
Table 2 Voxel alignment extracted features

| Voxel alignment feature number | Feature name               |
|-------------------------------|-----------------------------|
| 1                             | Short run emphasis          |
| 2                             | Long run emphasis           |
| 3                             | Intensity variability       |
| 4                             | Run-length variability      |
| 5                             | Run percentage              |
| 6                             | Low-intensity run emphasis  |
| 7                             | High-intensity run emphasis |
| 8                             | Low-intensity short-run emphasis |
| 9                             | High-intensity short-run emphasis |
| 10                            | Low-intensity long-run emphasis |
| 11                            | High-intensity long-run emphasis |

Table 3 Intensity-size-zone extracted features

| Intensity-size-zone feature number | Feature name                        |
|-----------------------------------|-------------------------------------|
| 1                                 | Short-zone emphasis                 |
| 2                                 | Large-zone emphasis                 |
| 3                                 | Intensity variability               |
| 4                                 | Size-zone variability               |
| 5                                 | Zone percentage                     |
| 6                                 | Low-intensity zone emphasis         |
| 7                                 | High-intensity zone emphasis        |
| 8                                 | Low-intensity short-zone emphasis   |
| 9                                 | High-intensity short-zone emphasis  |
| 10                                | Low-intensity large-zone emphasis   |
| 11                                | High-intensity large-zone emphasis  |
Table 4 ROC analysis test performance characteristics

| Parameter | ROC AUC | ROC threshold | ROC p   | Sensitivity | Specificity |
|-----------|---------|---------------|---------|-------------|-------------|
| R\textsubscript{NSR} | 88±5%   | >2.7%         | <0.0001 | 83%         | 83%         |
| R\textsubscript{Contrast} | 88±5%   | >4.7%         | <0.0001 | 83%         | 83%         |
| F1        | 78±7%   | >0.40         | 0.0001  | 89%         | 61%*        |
| F2        | 76±8%   | >0.077        | 0.001   | 100%        | 48%*        |
| F3        | 78±7%   | >0.34         | 0.001   | 94%         | 52%*        |

* p < 0.05 versus R\textsubscript{NSR}

F1 = Voxel-alignment high-intensity short-run emphasis; F2 = Voxel-alignment low-intensity long-run emphasis; F3 = Intensity-size-zone high-intensity short-zone emphasis.

Table 5 Logistic regression test performance characteristics

| Parameter | Logistic regression $\chi^2$ | Logistic regression p | % cases correctly predicted |
|-----------|-----------------------------|-----------------------|----------------------------|
| R\textsubscript{NSR} | 21.8                        | <0.0001               | 83%                        |
| R\textsubscript{Contrast} | 21.2                        | <0.0001               | 80%                        |
| F1        | 10.1                        | 0.002                 | 71%                        |
| F2        | 11.0                        | 0.001                 | 71%                        |
| F3        | 9.9                         | 0.002                 | 71%                        |

F1 = Voxel-alignment high-intensity short-run emphasis; F2 = Voxel-alignment low-intensity long-run emphasis; F3 = Intensity-size-zone high-intensity short-zone emphasis.

Table 6 ANOVA test performance characteristics
| Parameter    | F-ratio | ANOVA p | Negative     | Positive     |
|--------------|---------|---------|--------------|--------------|
| $R_{NSR}$    | 26.3    | <0.001  | 2.1±0.7%     | 3.3±0.7%     |
| $R_{Contrast}$ | 24.7    | <0.001  | 3.5±1.3%     | 5.7±1.5%     |
| F1           | 10.9    | 0.002   | 0.37±0.24    | 0.58±0.16    |
| F2           | 10.6    | 0.002   | 0.39±0.17    | 0.54±0.10    |
| F3           | 10.8    | 0.002   | 0.36±0.23    | 0.57±0.17    |

F1 = Voxel-alignment high-intensity short-run emphasis; F2 = Voxel-alignment low-intensity long-run emphasis; F3 = Intensity-size-zone high-intensity short-zone emphasis