Received: 20 July 2021 | Revised: 31 January 2022 | Accepted: 2 February 2022

DOI: 10.1002/mp.15537

RESEARCH ARTICLE

MEDICAL PHYSICS

Design and evaluation of a modular multimodality imaging phantom to simulate heterogeneous uptake and enhancement patterns for radiomic quantification in hybrid imaging: A feasibility study

Gijsbert M. Kalisvaart[1] | **Floris H.P. van Velden**[1] | **Irene Hernández-Girón**[1] | **Karin M. Meijer**[1] | **Laura M.H. Ghesquiere-Dierickx**[1,2,3] | **Wyger M. Brink**[1] | **Andrew Webb**[1] | **Lioe-Fee de Geus-Oei**[1,4] | **Cornelis H. Slump**[5] | **Dimitri V. Kuznetsov**[6] | **Dennis R. Schaart**[7,8] | **Willem Grootjans**[1]

[1]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

[2]Mechanical, Maritime, and Materials Engineering, Delft University of Technology, Delft, The Netherlands

[3]Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[4]Biomedical Photonic Imaging Group, University of Twente, Enschede, The Netherlands

[5]Robotics and Mechatronics, University of Twente, Enschede, The Netherlands

[6]Electronic and Mechanical Support Division, Delft University of Technology, Delft, The Netherlands

[7]Radiation Science and Technology, Delft University of Technology, Delft, The Netherlands

[8]Holland Proton Therapy Center, Delft, The Netherlands

Correspondence
Gijsbert M. Kalisvaart, Department of Radiology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands.
Email: g.m.kalisvaart@lumc.nl

Abstract

Background: Accuracy and precision assessment in radiomic features is important for the determination of their potential to characterize cancer lesions. In this regard, simulation of different imaging conditions using specialized phantoms is increasingly being investigated. In this study, the design and evaluation of a modular multimodality imaging phantom to simulate heterogeneous uptake and enhancement patterns for radiomics quantification in hybrid imaging is presented.

Methods: A modular multimodality imaging phantom was constructed that could simulate different patterns of heterogeneous uptake and enhancement patterns in positron emission tomography (PET), single-photon emission computed tomography (SPECT), computed tomography (CT), and magnetic resonance (MR) imaging. The phantom was designed to be used as an insert in the standard NEMA-NU2 IEC body phantom casing. The entire phantom insert is composed of three segments, each containing three separately fillable compartments. The fillable compartments between segments had different sizes in order to simulate heterogeneous patterns at different spatial scales. The compartments were separately filled with different ratios of ^{99m}Tc-pertechnetate, ^{18}F-fluorodeoxyglucose ([^{18}F]FDG), iodine- and gadolinium-based contrast agents for SPECT, PET, CT, and T_1-weighted MR imaging respectively. Image acquisition was performed using standard oncological protocols on all modalities and repeated five times for repeatability assessment. A total of 93 radiomic features were calculated. Variability was assessed by determining the coefficient of quartile variation (CQV) of the features. Comparison of feature repeatability at different modalities and spatial scales was performed using Kruskal-Wallis-, Mann-Whitney *U*-, one-way ANOVA- and independent *t*-tests.

Results: Heterogeneous uptake and enhancement could be simulated on all four imaging modalities. Radiomic features in SPECT were significantly less stable than in all other modalities. Features in PET were significantly less stable than in MR and CT. A total of 20 features, particularly in the gray-level

co-occurrence matrix (GLCM) and gray-level run-length matrix (GLRLM) class, were found to be relatively stable in all four modalities for all three spatial scales of heterogeneous patterns (with CQV < 10%).
Conclusion: The phantom was suitable for simulating heterogeneous uptake and enhancement patterns in [^{18}F]FDG-PET, ^{99m}Tc-SPECT, CT, and T_1-weighted MR images. The results of this work indicate that the phantom might be useful for the further development and optimization of imaging protocols for radiomic quantification in hybrid imaging modalities.

KEYWORDS
3D printing, hybrid imaging, multimodality imaging, phantom studies, radiomics, repeatability

1 | INTRODUCTION

Medical imaging has a pivotal role in personalizing the clinical management of cancer patients. With the ability to non-invasively quantify a myriad of physical, physiological, and molecular processes, the use of imaging techniques is an important prerequisite for adequate clinical staging, therapy response monitoring, and follow-up. In particular nuclear medicine imaging, including positron emission tomography (PET) and single-photon emission computed tomography (SPECT), has the ability to quantify and characterize cancer lesions with high precision.[1] Indeed, there has been much effort in recent years to develop and validate new quantitative image descriptors, also known as radiomic features, that capture the spatial distribution of uptake in cancer lesions.[2] Quantification of lesion texture and shape has been shown to provide important information for identifying specific tumor phenotypes, prediction of treatment resistance, and patient outcome.[3,4] Moreover, with the advent of hybrid imaging techniques, where X-ray computed tomography (CT) and magnetic resonance (MR) imaging are combined with PET and or SPECT, provides (near) simultaneous quantification of different physical and biological properties in a single imaging session and has extended the possibility to characterize cancer lesions with high precision.[4,5]

An important aspect of radiomic features is their accuracy and precision under different imaging conditions. In particular, the effects of image noise, reconstruction protocols, and motion artifacts are known to influence image quantification.[6–8] In addition to specific modality-dependent differences, such as spatial resolution, contrast resolution, and system sensitivity, technological evolutions such as improvements in detector technology or reconstruction algorithms may also influence image quantification.[8] Therefore, in order to accurately determine the value of radiomic features for characterizing cancer lesions in a clinical setting, it is of utmost importance to test the repeatability of these features and to know what lesion characteristics they can adequately capture under different imaging conditions.

Simulation of different imaging conditions can be achieved by performing standardized experiments using phantoms, allowing the determination of the precision or repeatability of radiomic features. Currently, a number of different phantom designs have been reported in the literature to test radiomic features under different imaging conditions.[9–12] However, these phantoms are typically designed to test such features only in a single imaging modality. In this study, a multimodality imaging phantom allowing testing of features in PET, SPECT, CT, and MR imaging was designed. Furthermore, the repeatability and the effect of spatial scale of heterogeneous uptake and enhancement patterns on features were investigated.

2 | MATERIALS AND METHODS

2.1 | Phantom design and construction

Several prototypes of the phantom were created in multiple design iterations to fulfill the predefined design criteria. Firstly, the phantom should fit within the casing of the National Electrical Manufacturers Association (NEMA-NU2) International Electrotechnical Commission (IEC) image quality (IQ) body phantom. Furthermore, the phantom should allow cross-modality imaging in the most commonly used tomographic imaging modalities, including PET, SPECT, CT, and MR imaging. Therefore, all phantom materials should be suitable for imaging in these modalities and the phantom should be re-fillable and re-usable. Additionally, simulation of heterogeneous uptake and enhancement patterns should be standardizable and reproducible at different spatial scales (in the range of several millimeters) in a single imaging session.

To this end, a modular design was used where the phantom consists of three cylindrical segments, containing different inserts, that can be interlocked and stacked to fit in the NEMA-NU2 IEC IQ casing (Figure 1). To create the segments, a polymethylmethacrylate tube (wall thickness 2 mm) with a diameter of 80 mm was cut to a length of 70 mm and the walls were routed down by means of computer numerical controlled milling, to 1 mm width at 10 mm from the edge of the tube. The segments were closed on each side by circular

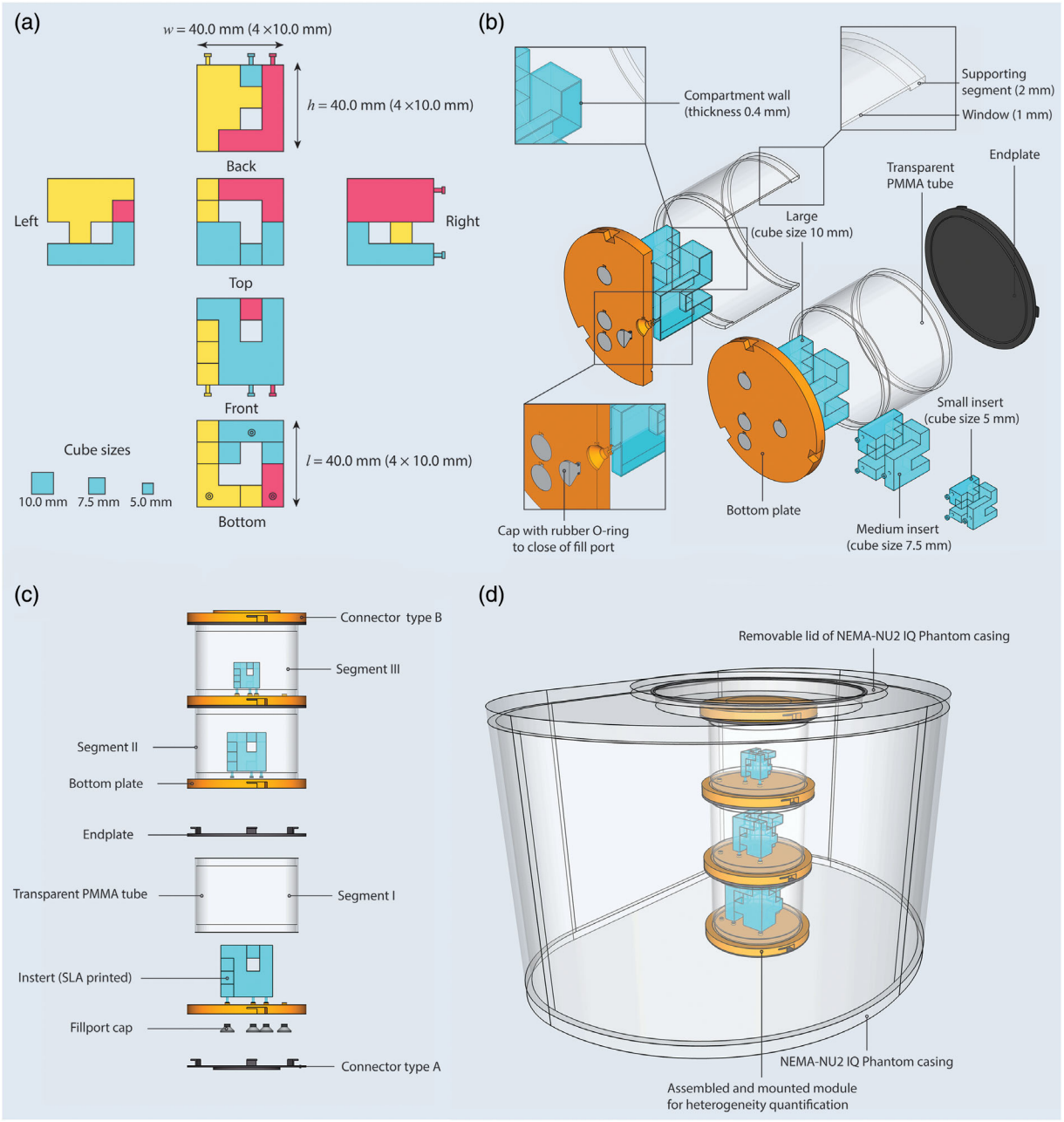

FIGURE 1 Schematic overview of a phantom segment. (a) Details regarding the design of the heterogeneity insert. The insert consists of three compartments (L- (red), T- (yellow), and U- (blue) shape) combined into a single cubic insert. The compartments are printed in three different sizes with an elemental cube size of 10.0 mm (large-sized), 7.5 mm (medium-sized), and 5.0 mm (small-sized). Different views of the insert are displayed for the largest heterogeneity insert in this figure (total dimensions 40.0 × 40.0 × 40.0 mm^3). (b) Details of an assembled phantom insert. The bottom plate (orange) contains ports to fill the compartments separately. The insert itself is enclosed by a polymethylmethacrylate (PMMA) tube and closed with an endplate (black). (c) Three segments are interlocked and combined into a single insert. Terminal connector plates (types A and B) are used to lock the assembled module in the circular cutouts (normally used to hold the lung insert of the original phantom) of the NEMA-NU2 IQ phantom casing. (d) Assembled phantom insert in the NEMA-NU2 IQ casing

endplates with a mechanical interlocking system. The plates were created by means of fused deposition modeling three-dimensional (3D) printing with polylactic acid.

Within each segment, an insert comprising three fillable compartments was positioned (L-, T- and U-shaped). The compartments were created using stereolithography (SLA) 3D printing and had a wall thickness of 0.4 mm. The design of the compartments was inspired by the digital phantom that was applied in the image biomarker standardization initiative (IBSI),[13,14] with some adjustments to ensure that the compartments could be created with currently available 3D printing techniques. The three compartments were based on geometric shapes composed of interconnected cubes

TABLE 1 Concentration of gadolinium-based contrast (Dotarem 0.5 millimoles per millilitre) and iodinated (Xenetix 350 mg I/ml) and activity concentrations [18F]-fluorodeoxyglucose (FDG) and sodium 99mTc-pertechnetate (Na[99mTcO4]) used during the phantom experiments. Target ratios between the different compartments with respect to the background are listed in the second column

	Ratio	MR [mmol/ml]	CT [mg I/ml]	PET [KBq/ml]	SPECT [KBq/ml]
Background body phantom	1	0.10×10^{-3}	0.7	2.1	3.6
Cylindrical insert	2	0.21×10^{-3}	1.3	4.1	5.9
L-shape	4	0.41×10^{-3}	2.6	8.2	12.2
T-shape	8	0.82×10^{-3}	5.2	16.5	24.4
U-shape	16	1.65×10^{-3}	10.4	33.0	48.8

Abbreviations: CT, computed tomography; MR, magnetic resonance; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

and were printed in one piece, forming a single cubic-shaped insert. Each compartment of the inserts was accessible through a separate fill-port in the bottom plate of the segment. The insert was printed in three different sizes by using a base cube size of 10, 7.5, and 5 mm, referred to as large-sized, medium-sized, and small-sized, respectively. After printing, leakage tests were conducted by means of a bubble test, as defined in ISO 20484:2017. During the test, the inserts were immersed in water at room temperature and kept at atmospheric pressure for 72 h. The test setup was monitored frequently to see any escape of bubbles that would indicate a leakage.

After assembling the three imaging segments of the phantom, the entire assembled module was placed in the NEMA-NU2 IQ casing by removing the top lid and placing it in the circular cut-out for the lung insert in the bottom plate. The phantom was then secured in the casing by replacing the top lid of the casing. High-resolution CT imaging was performed to determine the structural conformity of the printed compartments. Assessment of structural conformity was performed by performing unidimensional measurements of the dimensions in the IDS7 viewer (Sectra, Linköping, Sweden).

2.2 | Phantom preparation and image acquisition

Imaging experiments were conducted separately for MR, CT, PET, and SPECT by filling the compartments of the phantom with different concentrations of gadolinium-based contrast, iodinated contrast, [^{18}F]-fluorodeoxyglucose (FDG), or sodium ^{99m}Tc-pertechnetate (Na[^{99m}TcO4]), respectively. The concentration of the used substances was derived from clinical oncological imaging protocols.[15,16] For PET imaging, FDG concentrations were based on protocols defined in the research4life (EARL) FDG accreditation program.[17] For SPECT imaging, concentrations are based on a breast cancer protocol.[18] The ratios between concentrations were fixed for the background (1), cylindrical insert (2), L- (4), T- (8), and U-shaped (16). Details on the concentrations used during the experiments are summarized per modality in Table 1. For each modality, image acquisition was performed five times to evaluate the repeatability of the radiomic features. After each acquisition, the phantom was randomly repositioned (range of rotation, 1–20° and translation, 1–5 cm) in order to simulate patient repositioning. Between acquisition sessions, the compartments were emptied, flushed with water, emptied again, and air-dried for a few hours.

2.3 | Positron emission tomography

FDG-PET images were acquired on a Vereos PET/CT scanner (Philips Medical Systems, Best, The Netherlands) using two-bed positions with an acquisition time of 3 min per bed position. For each repeated scan, the time per bed position was increased to assure similar count statistics for each scan. The scanner was EARL accredited and reconstruction was performed using an EARL-based reconstruction protocol using a blob-based 3D iterative reconstruction algorithm (blobTOF; three iterations and 15 subsets) followed by a 5.5 mm full width at half maximum post-reconstruction Gaussian filter.[17] The image voxel size was $4 \times 4 \times 4$ mm^3. A low dose CT scan (40 mAs, 120 kVp) was acquired prior to the PET acquisition for the purpose of attenuation correction.

2.4 | Single-photon emission CT

SPECT images were acquired on a Discovery NM/CT 670 Pro (GE Healthcare, Chicago, Illinois, USA) using a low-energy, high-resolution collimator, noncircular orbit, step-and-shoot mode, 128 views (64 per camera head), and 20 s per view. Image reconstruction was performed using Evolution (GE Healthcare, Chicago, Illinois, USA), an ordered subsets expectation maximization (OSEM) algorithm incorporating collimator–detector response, attenuation and scatter correction, as well as resolution recovery. Images were reconstructed with nine iterations and 10 subsets, and a 128×128 matrices

(voxel size of 4.42 × 4.42 × 4.42 mm^3).[18] After reconstruction, the Q.Metrix software package (GE Healthcare, Chicago, Illinois, USA), automatically resampled both the CT and the SPECT images to a voxel size of 2.21 × 2.21 × 2.21 mm^3 prior to delineation. After the SPECT acquisition, low-dose CT images were acquired (100 kVp; auto tube current modulation of 100 mA) for the purpose of attenuation correction.

2.5 | MR imaging

MR imaging was performed using a 1.5T Ingenia MR scanner (Philips Healthcare, Best, The Netherlands). The integrated RF body coil was used for transmission and a torso-sized RF array coil was used for reception. Transverse T_1-weighted images were acquired using a 3D spoiled gradient-echo sequence (TR/TE = 7.6/4.6 ms, flip angle = 10°, voxel size = 1.33 × 1.33 × 2 mm^3, field of view = 320 × 320 × 240 mm^3, receiver bandwidth = 271 Hz/pixel) with an acquisition time of 3 min and 39 s. Images were reconstructed using vendor-supplied routines for image-based intensity normalization and three-dimensional gradient nonlinearity correction.

2.6 | Computed tomography

CT imaging was performed using an Aquilion ONE GENESIS Edition scanner (Canon Medical Systems Corporation, Otawara, Tochigi, Japan). An abdominal protocol was selected (liver, three phases, only portal phase used), with automatic exposure control on (so the system selects the mA depending on the phantom size and attenuation, SD10-Quality 5 mm for FC18 and AIDR3De), 80 × 0.5 mm^2 collimation, 120 kV, 0.813 pitch, 0.5 s rotation time, 139 mA average and $CTDI_{vol}$ = 4.4 mGy. Images were reconstructed with a FOV = 400.39 mm, 1 mm slice thickness and spacing (voxel size = 0.782 × 0.782 × 1 mm^3), using Adaptive Iterative Dose Reduction Enhanced (AIDR3De STD) as the reconstruction method and FC08 reconstruction kernel.

2.7 | Image analysis and feature selection

After image acquisition, images were cropped and registered using 3D Slicer (version 4.10.2; http://www.slicer.org) and Imalytics (version 3.2; Philips Research, Aachen, Germany) to a single reference image (in this case the high-resolution CT image of an empty phantom containing air). Segmentation was performed in 3D Slicer by defining a cylindrical-shaped volume of interest (VOI) for each segment separately with fixed dimensions (Ø, 60 mm and height, 50 mm) in order to eliminate the effect of modality-specific image segmentation results on radiomics calculation. Then, a segmentation mask was defined and exported together with the cropped images. The segmentation mask and images were loaded into PyRadiomics (version 3.0), and 107 features were calculated in accordance with the recommendations of the IBSI.[13] To enable the comparison of results derived from various imaging modalities, a fixed number of bins of 64 was used for all features and modalities (IBSI identifier: K15C). Since the same VOI was used for all images shape features (IBSI identifier: HCUG) were excluded from the analysis. This resulted in 93 features selected for the analyses, including 18 first order (IBSI identifier: UHIW and ZVCW), 24 gray-level co-occurrence matrix (GLCM, IBSI identifier: LFYI), 16 gray-level run-length matrix (GLRLM, IBSI identifier: TP0I), 16 gray-level zone size matrix (GLSZM, IBSI identifier: 9SAK), 14 gray-level dependence matrix (GLDM, IBSI identifier: REKO), and five neighboring gray-tone difference matrix (NGTDM, IBSI identifier: IPET)-features. First order features were calculated over the volume, GLCM and GLRLM features were averaged over 3D directions and GLSZM, GLDM, and NGTDM features were calculated from a single 3D matrix (IBSI identifiers: DHQ4, ITBB, and KOBO respectively). Symmetrical co-occurrence matrices and the Chebyshev norm with distance 1 (IBSI identifier: PVMT) were used for specific feature classes. No distance weighting was performed and a dependence coarseness value of 0 was used for GLDM features.

2.8 | Statistical analysis

Data analysis was divided into three sections; first, repeatability of radiomics compared between modalities per feature class was analyzed, then, repeatability between different insert sizes per modality was determined, and finally, the similarity of values between modalities per feature class was compared.

For the repeatability analysis (first and second analyses) the skewness of distributions was calculated for each radiomic feature per insert size and modality. Subsequently, repeatability of all radiomic features was expressed as the coefficient of quartile variation (CQV = (Q3 − Q1)/(Q1 + Q3) × 100%, where Q1 is the first quartile and Q3 is the third quartile), since the CQV is a relatively robust measure of dispersion in non-normal distributions.[19,20] Clinical test-retest studies typically report changes around 20% in standardized uptake values in tumors measured on [^{18}F]FDG-PET.[21] A threshold of CQV < 10% was chosen to identify features that are considered stable, whereas this threshold approaches differences of < 20% in values between Q1 and Q3. Groups of CQVs were tested on the homogeneity of variances with Levene's test before the

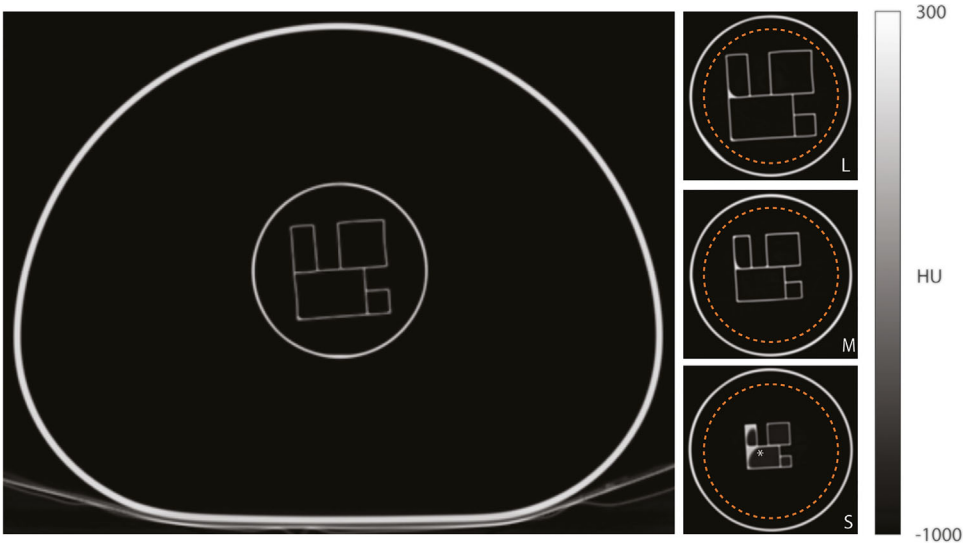

FIGURE 2 High-resolution computed tomography (CT) images of the heterogeneity phantom. The orange dashed lines represent the volume of interests (VOIs) of the detailed images of the large (L), medium (M), and small (S) inserts. Note that in the small-sized insert, some residual water from the leakage test is present in this image(*)

significance of differences of CQV values between groups was determined. When the hypothesis of homogeneity of variances was rejected ($p < 0.05$) non-parametric tests were used (Kruskal Wallis and Mann-Whitney *U* tests), whereas parametric tests were used otherwise (one-way ANOVA and independent *t*-test). The Kruskal Wallis or one-way ANOVA test was used to compare CQVs between the four different modalities (first section of the statistical analysis) and between the three different insert sizes (second section of the statistical analysis). When a significant difference was found the Mann-Whitney *U*-test or independent *t*-test was used for pairwise comparison between modalities or insert sizes. Bonferroni adjustment was applied to correct for multiple testing and the threshold for statistical significance was set at $p < p_{critical} = 0.005$. For comparison of feature values (third section of the statistical analysis), all individual radiomic values (X_i) were scaled to the same axis by dividing X_i to the mean of the feature for all modalities and spatial scales (X_m), i.e. $X_s = X_i/X_m$. For comparison of feature values, descriptive statistics were used. In this study, median values are reported with the corresponding first and third quartiles between parenthesis. Statistical analysis was performed using R (version 3.6.2; R Foundation for Statistical Computing, Vienna, Austria).

3 | RESULTS

3.1 | Visual assessment

High-resolution CT imaging showed that the compartments of the phantom were highly conformal and no large deviations with respect to shape or wall thickness were noticed (Figure 2). Furthermore, visual inspection of the PET, MR, and CT images showed that different structures could be visualized (Figure 3). Although PET was able to visualize structures of the medium-sized insert, the L-shaped compartment (contrast ratio 1:4) is barely distinguishable from the background. For the small-sized insert, only the U-shaped compartment (contrast ratio 1:16) could be readily visualized. For SPECT imaging, the U-shaped compartment for the large-sized and medium-sized inserts could be visualized, though the shape of the compartments is not adequately depicted in these images. The other compartments, could not be visualized under these conditions due to noise and partial volume effects. During data analysis, two small air bubbles were noted in the CT images of the large and small-sized cubic inserts (not visible in Figure 3). Voxels constituting the air bubbles in the VOI were masked. Overall, the repeatability analysis was not affected by masking these two air bubbles. Therefore, the results for the CT shown in this paper are based on the VOIs where the air bubbles are masked.

3.2 | Repeatability per modality and feature class

For all radiomic features, the median skewness per insert size and modality was 0.8 (0.4–1.5), suggesting that data were moderately skewed. The median CQV of all 93 radiomic features and spatial scales for MR, CT, PET and SPECT was 3.7% (1.7%–7.3%), 4.0% (1.4%–8.6%), 6.0% (2.8%–10.1%) and 17.7% (8.5%–27.0%), respectively (Table 2, and Table S1). A significant

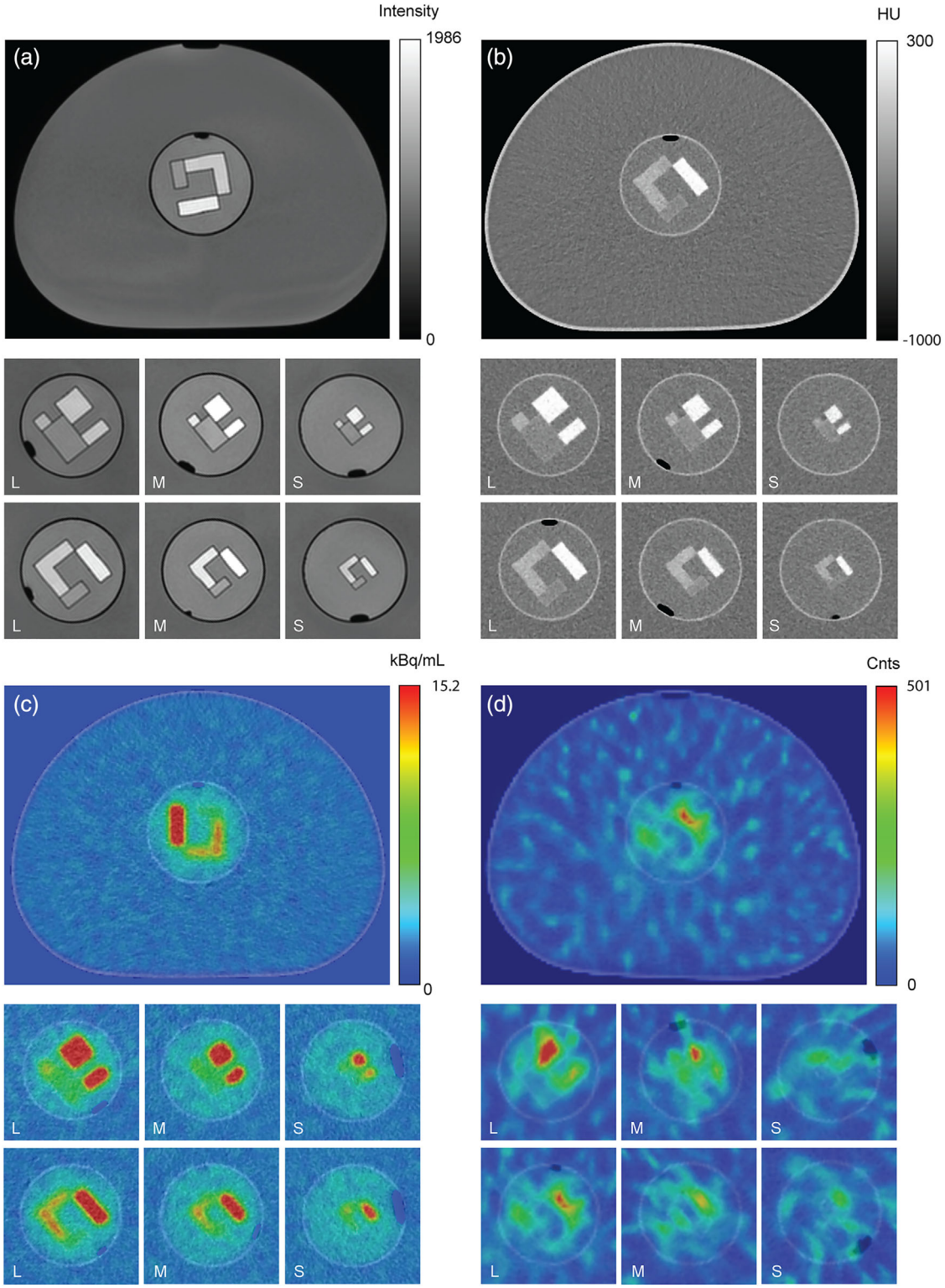

FIGURE 3 Results of the imaging experiments performed with the heterogeneity phantom on magnetic resonance (MR) (a), computed tomography (CT) (b), positron emission tomography (PET) (c), and single-photon emission computed tomography (SPECT) (d). Detailed images of the large (L), medium (M), and small (S) inserts, all aligned to the same orientation, are shown below their respective overview images. An air bubble is left to homogenize the solution in the background compartment, by giving it slightly a stir, just before the imaging experiments. The air bubble is excluded from the volume of interest (VOI) and did not impact radiomic feature quantification

TABLE 2 *p*-values for tests of significance of difference in CQV (%) values between modalities. $p < p_{critical} = 0.005$ is considered statistically significant, highlighted in blue. Corresponding *p*-values for Levene's test of homogeneity of variances are shown in parenthesis (Levene's test hypothesis was rejected when $p < 0.05$)

Class	All modalities	Modality	Median CQV	CT	PET	SPECT
All classes	<0.001(<0.001)	MR	3.70	0.2(<0.001)	<0.001(<0.001)	<0.001(<0.001)
		CT	3.99		<0.001(0.02)	<0.001(<0.001)
		PET	6.03			<0.001(<0.001)
		SPECT	17.68			
First order	<0.001(<0.001)	MR	1.95	0.2(0.8)	<0.001(0.6)	<0.001(<0.001)
		CT	1.24		<0.001(0.02)	<0.001(<0.001)
		PET	8.75			<0.001(<0.001)
		SPECT	16.74			
GLCM	<0.001(<0.001)	MR	2.60	0.2(0.2)	0.4(0.2)	<0.001(<0.001)
		CT	3.99		0.8(0.6)	<0.001(<0.001)
		PET	2.93			<0.001(<0.001)
		SPECT	12.05			
GLRLM	<0.001(<0.001)	MR	4.69	0.9(0.2)	0.7(<0.001)	<0.001(<0.001)
		CT	3.52		0.2(0.02)	<0.001(<0.001)
		PET	4.88			0.003(<0.001)
		SPECT	21.39			
GLSZM	<0.001(<0.001)	MR	4.98	0.005(<0.001)	0.01(<0.001)	<0.001(<0.001)
		CT	8.56		1.0(1.0)	0.01(0.9)
		PET	6.43			0.01(0.9)
		SPECT	19.48			
GLDM	<0.001(<0.001)	MR	5.21	0.04(0.1)	0.003(<0.001)	<0.001(<0.001)
		CT	7.18		0.1(<0.001)	<0.001(<0.001)
		PET	8.22			<0.001(0.5)
		SPECT	20.93			
NGTDM	<0.001(<0.001)	MR	6.67	0.6(0.3)	0.4(0.2)	<0.001(<0.001)
		CT	6.40		0.6(0.5)	<0.001(0.001)
		PET	5.99			0.001(0.01)
		SPECT	19.11			

Abbreviations: CT, computed tomography; GLCM, gray-level co-occurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run-length matrix; GLSZM, gray-level zone size matrix; MR, magnetic resonance; NGTDM, neighboring gray-tone difference matrix; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

difference in CQVs between different modalities was found ($p < 0.001$) (Figure 4). The CQVs in SPECT were significantly larger than the CQVs in MR, CT, and PET ($p < 0.001$ for all three). The CQVs in PET were significantly larger than in MR and CT ($p < 0.001$ for both). In class-specific analyses, SPECT radiomic features were found to be less repeatable in all classes and modalities (except the GLSZM class in PET and CT images), while PET features were only less repeatable than CT features of the first order class ($p < 0.001$) and MR features of the first order and GLDM classes ($p < 0.001$ and 0.003, respectively). No significant differences in repeatability were found when comparing MR and CT.

3.3 | Feature repeatability for different spatial scales

Overall, 20 radiomic features were found to be stable with CQVs < 10% for all spatial scales and imaging modalities (Table S1), of which 8 and 6 features from the GLCM and GLRLM class, respectively. In MR, CT, PET, and SPECT, 77, 58, 66, and 21 features were stable with a CV < 10% for all insert sizes. The CQVs differed significantly per insert size in both MR and CT images ($p < 0.001$ for both) (Table 3). The CQVs in MR images were significantly lower in the large-sized insert versus the medium- and small-sized inserts ($p < 0.001$ for both) while no significant difference in CVs was found

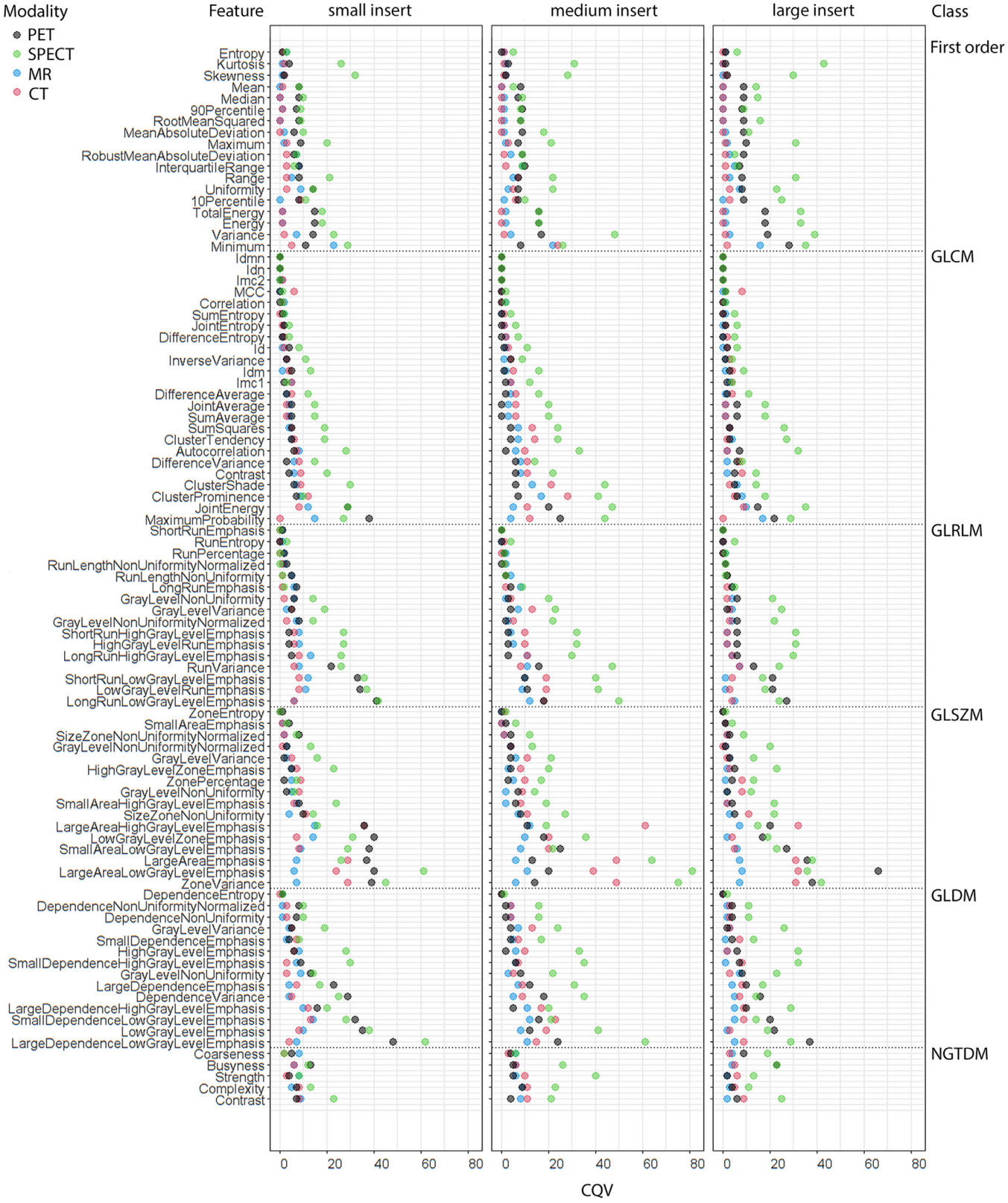

FIGURE 4 Scatter plot depicting the CQVs (%) per feature, insert size, and modality. On the y-axis, features are ordered, per class, from the lowest to highest median CQV over all modalities and insert sizes

TABLE 3 *p*-values for tests of significance of difference in CQV (%) values between insert sizes per modality. $p < p_{critical} = 0.005$ is considered statistically significant, indicated by the blue cell color. Corresponding *p*-values for Levene's test of homogeneity of variances are shown in parenthesis (Levene's test hypothesis was rejected when $p < 0.05$)

Modality	All insert sizes	Insert size	Median CQV	Medium	Large
MR	<0.001(0.003)	small	5.35	0.7(0.8)	<0.001(0.001)
		medium	4.72		<0.001(<0.001)
		large	2.35		
CT	<0.001(<0.001)	small	4.00	0.03(<0.001)	0.3(1.0)
		medium	6.40		<0.001(0.001)
		large	3.06		
PET	0.09(<0.001)	small	6.16	–	–
		medium	4.94		–
		large	6.09		
SPECT	0.02(0.05)	small	14.95	–	–
		medium	20.70		–
		large	17.83		

Abbreviations: CT, computed tomography; MR, magnetic resonance; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

between the medium- and small-sized inserts ($p = 0.7$). In CT, measurements on the medium-sized insert were significantly less repeatable than measurements on the large-sized inserts ($p < 0.001$), while no differences were found comparing the large- and medium-sized to the small-sized insert ($p = 0.03$ and 0.03, respectively). PET and SPECT CQVs did not show a significant correlation with spatial scale ($p = 0.09$ and $p = 0.02$, respectively).

3.4 | Similarity per feature class

The largest differences in radiomic feature values between modalities were found in the first order class (Figure 5). In descending order median X_s for first order class features were 3.0 (1.1–3.9), 0.5 (0.2–0.8), 0.06 (0.03–0.1), 0.04 (0.01–0.08) for PET, MR, SPECT, and CT, respectively. Features in the GLCM class showed high similarity between different imaging modalities, with a median X_s of 1.0 (0.8–1.4), 1.0 (0.8–1.2), 1.0 (0.7–1.1), 0.9 (0.6–1.0) in PET, MR, SPECT, and CT, respectively. Median values for all features are presented in the supplementary material, Table S2, for reference.

4 | DISCUSSION

In this study, a newly developed multimodality imaging phantom was proposed for the purpose of simulating heterogeneous uptake and enhancement patterns in the most common tomographic imaging modalities. Imaging experiments on PET, MR, CT, and SPECT systems showed that the phantom can be successfully used to simulate and quantify such patterns in a standardized fashion. Furthermore, the modular design allows multiple experiments to be performed in a single imaging session. Besides allowing a standardized approach for simulating heterogeneous uptake and enhancement patterns, it is reusable and fits within a standard NEMA-NU2 IQ phantom case, permitting the benchmarking of different imaging modalities.

Interest in the design of imaging phantoms that can simulate heterogeneous uptake and enhancement patterns in medical imaging has increased over the last few years. In particular, different research groups have found that the use of such imaging phantoms is important for assessing the quantitative accuracy of different imaging protocols.[22] In literature, several approaches have been described for creating suitable phantoms. These approaches can be categorized as phantoms that use materials that can be cast into molds, contain fillable compartments, or are formed layer-by-layer by printing materials compatible with the specific imaging modality being investigated.[23,24] For cast phantoms, reusability, and standardization for multiple imaging modalities are often challenging. This is due to the fact that materials (usually liquids) need to be set after casting, thereby expanding or shrinking (changing the geometry of structures). Furthermore, the shelf life of these phantoms is usually limited due to degradation of the used material and in the case of materials suitable for nuclear imaging techniques (PET and SPECT) half-life of the used radioactive isotopes. Furthermore, the geometry that can be used is limited and requires a careful assessment when assembling different structures seamlessly together. A layer-by-layer creation, by for example printing radioactive resin (SPECT and PET)

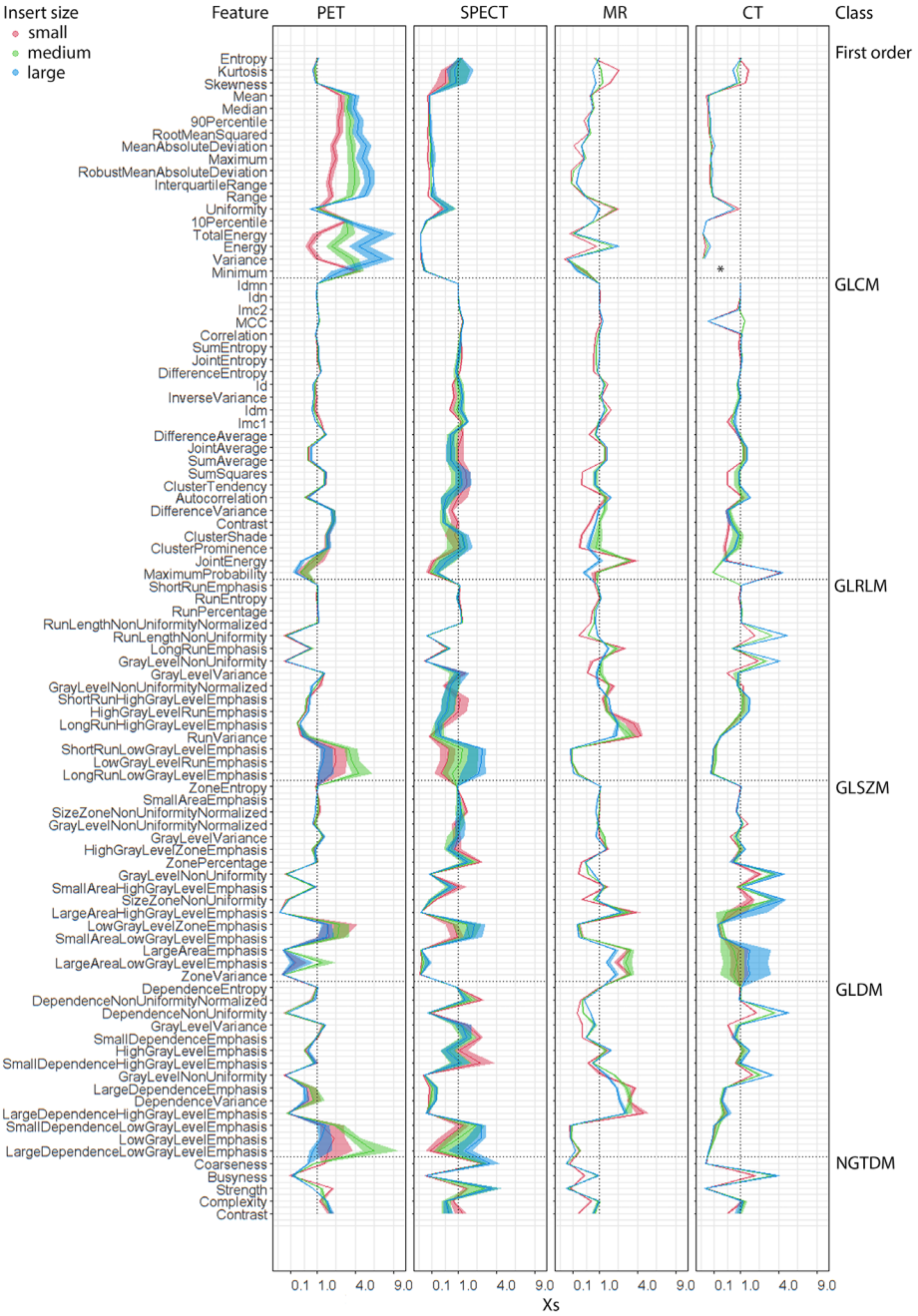

FIGURE 5 Line plots showing the median and range (shaded) of scaled values (Xs) of the radiomic features. The x-axis presents the Xs on a square root scale. Similar to Figure 4, the y-axis features are ordered, per class, from the lowest to highest median CQV over all modalities and insert sizes. *Since minimum voxel values in CT images were negative, this feature is not plotted on the square root scale for CT

or radio-opaque dyes (CT) on paper, is a convenient way to create realistic anthropomorphic phantoms.[25,26] However, the reusability of such phantoms for PET and SPECT is challenging due to the radioactive decay of the radioactive isotopes. Moreover, once the phantoms have been constructed using these approaches, contrast and geometry are fixed and cannot be varied easily.

Given the importance of standardization and multi-institutional comparison, this study focused on the creation of a phantom using separately-fillable compartments. Although an imaging phantom using compartments is a flexible and practical way of creating volumes with different image contrast, it also has limitations. These are related to the finite thickness of the walls surrounding the compartments and the ease of filling the compartments (particularly at small scales). Although the compartments designed in this phantom have been produced using high-resolution SLA printing techniques, the wall thickness could be reduced no further than 0.4 mm due to the physical constraints of the

printing technique itself. While no cold-wall effects were observed with PET and SPECT imaging, the walls were visualized on CT and MR images.[27] Further experimental work is required to determine whether the wall thickness can be reduced even further without compromising the structural integrity of the phantom itself. Continuous improvements in 3D printing methods and the development of new materials with different physical and chemical properties could help to overcome the current limitations. Although the current phantom has been tested for re-usability, more experiments are required to determine the accuracy of filling and re-filling of the compartments and the practicality of the use of such a phantom in a multicenter setting.

Another important aspect is the geometry of the fillable compartments which should be designed in such a way that different types of radiomic features, for example, shape or texture features, can be tested. This includes radiomic features that quantify heterogeneity at different levels, including global, regional, and voxel-to-voxel variations.[28] Although extensive study of the optimal geometrical configuration of different compartments was not the aim of this study, the compartments were designed such that these aspects of heterogeneity were represented for imaging with relatively low spatial resolution, such as PET and SPECT. As an alternative to a geometric phantom, as presented here, anthropomorphic phantoms can be used. Anthropomorphic phantoms are typically designed by segmenting organs or lesions from clinical images.[22] Although these phantoms mimic lesions in patients, there is a bias towards the use of practicable patient data to generalize the problem of radiomic quantification and not all features might be tested to their full extent. Furthermore, depending on the reconstruction and postprocessing algorithms used, artifacts and segmentation inaccuracies can result in significant deviations from the actual lesions present in the patient. Finally, the appearance of lesions is different on different imaging modalities, with different contrast and patterns of heterogeneity. For example, heterogeneity in FDG-uptake observed in PET does not necessarily correspond to the enhancement patterns observed in MR or CT. With a geometrical design, the shape of a phantom is mathematically determined and limited by the manufacturing techniques. The design can thus be altered to test limitations in modality-specific characteristics (such as spatial resolution, shape, and distribution of the patterns).

Although phantoms are useful for the optimization and harmonization of imaging protocols, other methods are also available for this purpose. One of such methods, designated COMBAT (Combing Batches), can be used to harmonize images using empirical Bayes methods.[29] If employed appropriately, this method can be used to perform multicenter comparisons without changing local imaging protocols. Although COMBAT can harmonize data from heterogeneous sources, it can lead to loss of physical meaning. Therefore, no direct method currently exists to apply previously determined harmonization transformation to radiomic features derived from a new patient in a different center.[30] Another potential alternative to phantoms is the use of advanced simulation software, such as GATE (geant4 application for tomographic emission), to simulate the characteristics of different scanners.[31] However, assessment of the actual performance of a specific on-site scanner (with specific non-idealities) is not directly possible using such methods. Furthermore, with many different scanners manufacturers and imaging protocols, simulation of all these different conditions can quickly become too complex and computationally expensive. Thus, the use of phantoms is the most direct way of assessing the performance of locally-installed imaging systems in combination with local imaging protocols.

Although the current results on provide insight into the relative stability of the radiomics features, several factors might have influenced the reported repeatability results. The radiomic features calculated on MR and CT might be relatively repeatable given that the size of the insert compartments is significantly larger than the spatial resolution and reconstructed voxel sizes of these modalities. The shape of the current phantom was designed to represent heterogeneity and test partial volume effects on imaging modalities with the lowest spatial resolution observed in hybrid imaging, i.e. PET and SPECT. In addition, further reduction of the insert sizes was limited by the physical restraints of the current printing techniques. With the improvement of printing techniques, efforts should be made to further decrease the spatial scale of these heterogeneity inserts and optimize the use of this phantom for repeatability testing in imaging modalities with a higher spatial resolution, i.e. MR and CT. Another factor influencing repeatability is phantom repositioning between imaging sessions. Although the phantom was rotated around and translated over the x and y axes, no z-axis tilt was performed during repositioning. This impairs the translatability of the current findings to clinical images since such a z-axis tilt might occur in a clinical setting when repositioning patients. Moreover, The effects of repositioning on both rotationally invariant and rotationally variant features should be studied in future research. Texture features require interpolation of anisotropic voxels to isotropic voxels to be rotationally invariant. No interpolation was performed in this study since interpolation affects obtained radiomic values and we aimed to resemble the way images are acquired and analyzed in clinical practice as much as possible. Therefore, the effect of phantom repositioning on feature repeatability is expected to be larger for MR and CT (anisotropic voxels) than for PET and SPECT (isotropic voxels) in this study.[32,33] Indeed, our results show that the differences in highly repeatable MR and CT features compared to less repeatable PET and SPECT features are less prominent in texture

features than in first order features (Table 2). Moreover, a harmonized data quantization method was used in this study to improve multimodality comparability. A fixed number of bins of 64 was chosen for the radiomics analysis for all modalities while alternative discretization settings might influence feature repeatability differently in specific modalities.[34,35] Future studies should aim to assess multimodality radiomic feature repeatability for various simulations of patient repositioning, and different image processing- and radiomics calculation settings for both rotationally invariant and rotationally variant features. Another area of interest for future studies is the definition of unequivocal ground truth to test the accuracy of absolute radiomics quantification. Additionally, phantom design (such as shape, size, and orientation of the fillable compartments) should be optimized to test different feature types (e.g., tumor-to-background ratios), that are often reported in clinical studies and provide important information.[36]

Results from this study show that, with the proposed phantom, heterogeneous uptake and enhancement patterns could be simulated on all four tomographic imaging modalities. Repeatability assessment showed that radiomic features derived from T_1-weighted MR and CT images generally had lower variability compared to PET and SPECT. This can be attributed to the lower noise levels and less pronounced partial volume effects in these images. Furthermore, radiomic features derived from SPECT images had the highest variability, and the different compartments of the phantom could also not be readily visualized in these images. The occurrence of artifacts and the relatively high levels of image noise in ^{99m}Tc-SPECT are known to hinder the quantification of radiomic features and result in increased variability.[18,37,38] Although multimodality phantom studies assessing repeatability lack in literature, the current findings outline variabilities of comparable magnitude as findings in other patient- and single-modality phantom studies.[7,11,12,39] While direct comparison of our results with patient data on a single feature level is not within the scope of this article, our findings correspond with the general consensus that specifically first order entropy is relatively stable between different imaging modalities in a clinical setting.[40] Moreover, the first order entropy values found in this study are in the same order of magnitude as those typically found in patient studies.[28]

5 | CONCLUSION

In this study, the design and first evaluation of a multimodality imaging phantom have been described. The proposed design permits the simulation of heterogeneous uptake and enhancement patterns in the most commonly used tomographic imaging modalities in hybrid imaging. Furthermore, repeatability assessment for radiomics was performed, showing that overall variability of radiomic features derived from T_1-weighted MR, CT, and higher order radiomic features derived from [^{18}F]FDG-PET images was acceptable under the tested imaging conditions. However, first order [^{18}F]FDG-PET features and all features derived from ^{99m}Tc-SPECT images showed larger variability. Future studies should address the reproducibility of radiomics quantification under varying imaging conditions in a multicenter setting and further evaluate the use of the proposed phantom for the standardization of imaging protocols across different imaging platforms.

ACKNOWLEDGMENTS

The authors would like to thank Nan Huang, Juliana Sabelis, Ernst van der Wal, Joeri Kuil, and Petra Dibbets-Schneider for their contribution to the study. Funding was provided by an educational grant from Philips Electronics Nederland B. V., Eindhoven, The Netherlands and a public grant from TKI Life Sciences & Health, Health~Holland.

CONFLICT OF INTEREST

We declare the following financial interests/personal relationships which may be considered as potential competing interests: Gijsbert M. Kalisvaart is the recipient of an educational grant from Philips Electronics Nederland B. V., Eindhoven, The Netherlands, during the writing of this manuscript. Furthermore, the research presented in the manuscript is supported by a public grant from TKI Life Sciences & Health, Health~Holland. No other potential conflicts of interest relevant to this article exist.

REFERENCES

1. Grootjans W, de Geus-Oei LF, Troost EG, Visser EP, Oyen WJ. Bussink J. PET in the management of locally advanced and metastatic NSCLC. *Nat Rev Clin Oncol*. 2015;12(7):395-407.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577.
3. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446.
4. Lucia F, Visvikis D, Vallieres M, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging*. 2019;46(4):864-877.
5. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60(14):5471-5496.
6. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-1248.
7. van Velden FH, Kramer GM, Frings V, et al. Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol Imaging Biol*. 2016;18(5):788-795.
8. Grootjans W, Tixier F, van der Vos CS, et al. The impact of optimal respiratory gating and image noise on evaluation of intratumor heterogeneity on 18F-FDG pet imaging of lung cancer. *J Nucl Med*. 2016;57(11):1692-1698.

9. Valladares A, Beyer T, Rausch I. Physical imaging phantoms for simulation of tumor heterogeneity in PET, CT, and MRI: an overview of existing designs. *Med Phys*. 2020;47(4):2023-2037.
10. Pfaehler E, Beukinga RJ, de Jong JR, et al. Repeatability of (18) F-FDG PET radiomic features: a phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med Phys*. 2019;46(2):665-678.
11. Baessler B, Weiss K, Pinto Dos Santos D. Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. *Invest Radiol*. 2019;54(4):221-228.
12. Pfaehler E, van Sluis J, Merema BBJ, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. *J Nucl Med*. 2020;61(3):469-476.
13. Zwanenburg A, Vallieres M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338.
14. Lambin P. Radiomics Digital Phantom, CancerData. 2016. DOI:10.17195/candat.2016.08.1.
15. Bellin MF. MR contrast agents, the old and the new. *Eur J Radiol*. 2006;60(3):314-323.
16. Bae KT. Intravenous contrast medium administration and scan timing at CT: considerations and approaches. *Radiology*. 2010;256(1):32-61.
17. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: eANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42(2):328-354.
18. Collarino A, Pereira Arias-Bouda LM, Valdes Olmos RA, et al. Experimental validation of absolute SPECT/CT quantification for response monitoring in breast cancer. *Med Phys*. 2018;45(5):2143-2153.
19. Bonett DG. Confidence interval for a coefficient of quartile variation. *Comput Stat Data An*. 2006;50(11):2953-2957.
20. Altunkaynak B, Gamgam H. Bootstrap confidence intervals for the coefficient of quartile variation. *Commun Stat-Simul C*. 2019;48(7):2138-2146.
21. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50(Suppl 1):122S-50S.
22. Gallivanone F, Interlenghi M, D'Ambrosio D, Trifiro G, Castiglioni I. Parameters influencing PET imaging features: a phantom study with irregular and heterogeneous synthetic lesions. *Contrast Media Mol Imaging*. 2018:5324517.
23. Filippou V, Tsoumpas C. Recent advances on the development of phantoms using 3D printing for imaging with CT, MRI, PET, SPECT, and ultrasound. *Med Phys*. 2018;45(9):740-760.
24. Keenan KE, Ainslie M, Barker AJ, et al. Quantitative magnetic resonance imaging phantoms: a review and the need for a system phantom. *Magn Reson Med*. 2018;79(1):48-61.
25. Jahnke P, Schwarz S, Ziegert M, Schwarz FB, Hamm B, Scheel M. Paper-based 3D printing of anthropomorphic CT phantoms: feasibility of two construction techniques. *Eur Radiol*. 2019;29(3):1384-1390.
26. Lappchen T, Meier LP, Furstner M, et al. 3D printing of radioactive phantoms for nuclear medicine imaging. *Eur J Nucl Med Mol Imaging Phys*. 2020;7(1):22.
27. Berthon B, Marshall C, Edwards A, Evans M, Spezi E. Influence of cold walls on PET image quantification and volume segmentation: a phantom study. *Med Phys*. 2013;40(8).
28. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med*. 2011;52(3):369-378.
29. WE Johnson, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127.
30. Da-Ano R, Masson I, Lucia F, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci Rep*. 2020;10(1):10248.
31. Cuplov V, Pain F, Jan S. Simulation of nanoparticle-mediated near-infrared thermal therapy using GATE. *Biomed Opt Express*. 2017;8(3):1665-1681.
32. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. *J Nucl Med*. 2015;56(11):1667-1673.
33. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062.
34. Leijenaar RT, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep*. 2015;5:11075.
35. Larue R, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol*. 2017;56(11):1544-1553.
36. Braman NM, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res*. 2017;19(1):57.
37. Grootjans W, Meeuwis AP, Slump CH, de Geus-Oei LF, Gotthardt M, Visser EP. Performance of 3DOSEM and MAP algorithms for reconstructing low count SPECT acquisitions. *Z Med Phys*. 2016;26(4):311-322.
38. Peters SMB, van der Werf NR, Segbers M, et al. Towards standardization of absolute SPECT/CT quantification: a multi-center and multi-vendor phantom study. *Eur J Nucl Med Mol Imaging Phys*. 2019;6(1):29.
39. Desseroit MC, Tixier F, Weber WA, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. *J Nucl Med*. 2017;58(3):406-411.
40. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1143-1158.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kalisvaart GM, van Velden FHP, Hernández-Girón I, et al. Design and evaluation of a modular multimodality imaging phantom to simulate heterogeneous uptake and enhancement patterns for radiomic quantification in hybrid imaging: A feasibility study. *Med Phys*. 2022;1-14. https://doi.org/10.1002/mp.15537