



Robustness of pulmonary nodule radiomic features on computed tomography as a function of varying radiation dose levels—a multi-dose *in vivo* patient study

Gijs A. Bartholomeus¹ · Wouter A. C. van Amsterdam¹ · Annemarie M. den Harder¹ · Martin J. Willemink² · Robbert W. van Hamersvelt¹ · Pim A. de Jong¹ · Tim Leiner^{1,3}

Received: 12 May 2022 / Revised: 16 March 2023 / Accepted: 28 March 2023 / Published online: 19 April 2023

© The Author(s) 2023

Abstract

Objective Analysis of textural features of pulmonary nodules in chest CT, also known as radiomics, has several potential clinical applications, such as diagnosis, prognostication, and treatment response monitoring. For clinical use, it is essential that these features provide robust measurements. Studies with phantoms and simulated lower dose levels have demonstrated that radiomic features can vary with different radiation dose levels. This study presents an *in vivo* stability analysis of radiomic features for pulmonary nodules against varying radiation dose levels.

Methods Nineteen patients with a total of thirty-five pulmonary nodules underwent four chest CT scans at different radiation dose levels (60, 33, 24, and 15 mAs) in a single session. The nodules were manually delineated. To assess the robustness of features, we calculated the intra-class correlation coefficient (ICC). To visualize the effect of milliampere-second variation on groups of features, a linear model was fitted to each feature. We calculated bias and calculated the R^2 value as a measure of goodness of fit.

Results A small minority of 15/100 (15%) radiomic features were considered stable ($ICC > 0.9$). Bias increased and R^2 decreased at lower dose, but shape features seemed to be more robust to milliampere-second variations than other feature classes.

Conclusion A large majority of pulmonary nodule radiomic features were not inherently robust to radiation dose level variations. For a subset of features, it was possible to correct this variability by a simple linear model. However, the correction became increasingly less accurate at lower radiation dose levels.

Clinical relevance statement Radiomic features provide a quantitative description of a tumor based on medical imaging such as computed tomography (CT). These features are potentially useful in several clinical tasks such as diagnosis, prognosis prediction, treatment effect monitoring, and treatment effect estimation.

Key Points

- The vast majority of commonly used radiomic features are strongly influenced by variations in radiation dose level.
- A small minority of radiomic features, notably the shape feature class, are robust against dose-level variations according to ICC calculations.
- A large subset of radiomic features can be corrected by a linear model taking into account only the radiation dose level.

Keywords Humans · Linear models · Tomography (x-ray computed) · Multiple pulmonary nodules · Radiation dosage

Abbreviations

ANOVA Analysis of variance
CCC Concordance correlation coefficient

COV Coefficient of variation
FBP Filtered back projection
GLCM Gray-level co-occurrence matrix
GLDM Gray-level dependence matrix
GLRM Gray-level run length matrix
GLSZM Gray-level size zone matrix
ICC Intra-class correlation coefficient
LoG Laplacian of Gaussian
MSE Mean square error
MSR Mean square for rows
ROI Region of interest

✉ Gijs A. Bartholomeus
Ga.bartholomeus@gmail.com

¹ University Medical Center Utrecht, Utrecht, the Netherlands

² Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

³ Mayo Clinic, Rochester, MN, USA

Introduction

Advances in data science have led to a surge in imaging biomarkers to improve lung cancer diagnosis, prognostication, and treatment response monitoring. Among these modern biomarkers is the class of computed tomography (CT) radiomic features. Radiomics is defined as the quantification of CT radiographic phenotype using data-characterization algorithms [1, 2]. Statistical models are used to relate these radiomic features to diagnosis, prognostication, and treatment response.

Early detection of possibly malignant pulmonary nodules would make it possible to start therapy in an earlier stage which is prognostically favorable [3]. Conversely, early discrimination of benign nodules from malignant nodules would relieve patients from unnecessary follow-up CT scans. Thus, the goal of radiomics is to go beyond morphological imaging and to aid in the diagnosis, classification, and therapeutic decision-making of patients who undergo radiographic imaging using statistical models.

For radiomic features to be useful in the clinical process, feature values need to be reproducible. This is to say, a feature should be influenced primarily by biological traits of the patient, and not by external conditions such as the type of CT equipment, reconstruction algorithm, region of interest (ROI) selection and segmentation, etc. A drawback of radiomic features is that they seem to be sensitive to conditions currently not standardized in clinical care. One of these variables in CT scanning is the tube current–time product (milliampereseconds, or mAs). Computed tomography is a major source of radiation exposure related to medical imaging. To reduce the dose, the level of milliampereseconds is lowered at the cost of increased image noise [4]. Because image noise increases non-linearly with decreasing milliampereseconds [5], we hypothesize that this increase in noise will influence radiomic feature values. Although some phantom studies have shown that the effect of varying tube current on radiomic features does not significantly affect radiomic features [6], other studies have shown milliamperesecond variation does in fact significantly influence radiomic feature values [7, 8]. Although several *in vivo* dose modulation radiomic feature robustness studies have been performed to date, these studies are retrospective in the sense that they compare features taken from a single diagnostic scan, and later follow-up scans [9, 10]. As mentioned in the systematic review by Reiazi et al: “The drawbacks of the retrospective studies are that the investigators did not have control over the parameters studied, and the range of the scan acquisition parameter variations were limited to those used in imaging patients.” [11]. Our study differs from these studies in that multiple scans with different radiation doses were obtained in a single examination within a time frame of approximately 20 min.

Therefore, we sought to investigate the *in vivo* robustness of pulmonary nodule radiomic features in patients who

underwent chest CT scans at four different radiation dose levels.

Methods

Study population and image acquisition

In this study, patients 50 years or older with 1 or more known pulmonary nodules scheduled for a follow-up chest CT were eligible for inclusion. Detailed inclusion criteria are listed in Appendix 1. IRB approval was given under reference number NL46146.041.13 [12, 13]. Participants signed a written informed consent form prior to inclusion in the study.

A 256-slice CT system (Brilliance iCT; Philips Healthcare) was used for image acquisition. Patients were asked to hold their breath at deep inspiration during each acquisition. After scout images were obtained, image acquisition was performed using our routine non-enhanced chest CT protocol, immediately followed by 3 acquisitions at reduced radiation dose levels. Automatic current selection was only used for the reference protocol and modified to the values as described for the lower-dose acquisitions. Z-axis dose modulation and dynamic angular dose modulation were not used to minimize variation.

All acquisitions were performed with the same length (Z coverage). Images were reconstructed with a slice thickness of 1 mm and an increment of 0.7 mm. Tube current–time products of 60 (reference dose), 33 (45% reduction), 24 (60% reduction), and 15 mAs (75% reduction) were used in combination with a tube voltage of 100 kV for patients with a weight less than 80 kg and a tube voltage of 120 kV for patients with a weight greater than 80 kg. Gantry rotation time was 0.33 s with a pitch of 0.758. No contrast medium was injected. Scans were reconstructed using filtered back projection (FBP). Data will be made available for non-commercial purposes upon reasonable request to the authors.

Segmentation

For the evaluation of the stability of radiomic features of pulmonary nodules on computed tomography, pulmonary nodules were manually segmented in the open-source image processing software platform 3D Slicer (Slicer.org). Nodules were independently identified by two experienced radiologists to make sure no pulmonary nodules were missed. For each scan, a binary (3D) label map annotating the pulmonary nodules for each radiation dose level was created by manual segmentation with the help of the semiautomatic “grow from seeds” region growing volumetric segmentation algorithm [14]. Contours were generated by one author (G.B.) and independently verified by an experienced radiologist (P.J.).

Radiomic features

The open-source python package for the extraction of radiomic features from medical imaging Pyradiomics (version 2.2.0) was used to extract the radiomic features [15]. Statistical analysis was done in R (version 4.10.2). Seven different filters were applied to the images before feature extraction (including original image, no filter). Per filter, 86 features were extracted, divided into six different feature classes. The following feature classes were extracted: shape (only for the original image); gray-level co-occurrence matrix (GLCM); gray-level dependence matrix (GLDM); first-order, gray-level run length matrix (GLRM); and gray-level size zone matrix (GLSZM) [15]. A detailed list of extracted features can be found in Appendix 2.

Statistical analysis

Statistical analysis was performed on a nodule level, using the package psych (version 1.9.11) in R. The intra-class correlation coefficient 3.1 (ICC) was calculated to assess feature robustness [16] by assessing agreement in radiomic feature values between CT scans acquired with different radiation doses, and is calculated as follows:

$$ICC = \frac{MS_R - MS_E}{MS_R + (k - 1)MS_E}$$

where MS_R = mean square for rows, MS_E = mean square error, and k = number of different radiation dose levels. According to Koo et al, ICC values less than 0.5 were

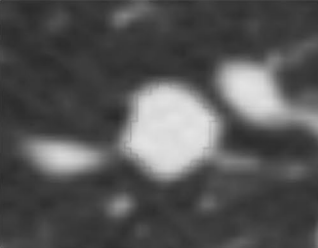
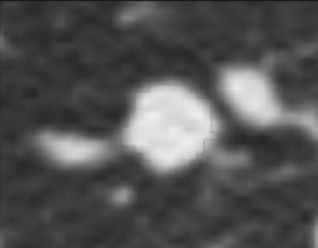




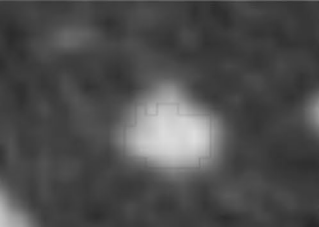

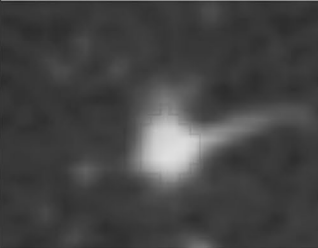
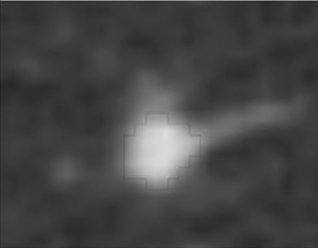


			
(Shape) Sphericity: 0.859 (GLRLM) Gray level non-uniformity: 79.915 (GLCM) Cluster Prominence: 17892.19	(Shape) Sphericity: 0.846 (-1.5%) (GLRLM) Gray level non-uniformity: 62.194 (-22%) (GLCM) Cluster Prominence: 49288.98 (175%)	(Shape) Sphericity: 0.851 (-0.9%) (GLRLM) Gray level non-uniformity: 61.135 (-23%) (GLCM) Cluster Prominence: 27304.78 (51%)	(Shape) Sphericity: 0.851 (-0.9%) (GLRLM) Gray level non-uniformity: 57.440 (-29%) (GLCM) Cluster Prominence: 37394.47 (108%)
			
(Shape) Sphericity: 0.799 (GLRLM) Gray level non-uniformity: 10.104 (GLCM) Cluster Prominence: 96685.23	(Shape) Sphericity: 0.816 (2.1%) (GLRLM) Gray level non-uniformity: 9.045 (-11%) (GLCM) Cluster Prominence: 21754.21 (78%)	(Shape) Sphericity: 0.813 (1.8%) (GLRLM) Gray level non-uniformity: 7.376 (-27%) (GLCM) Cluster Prominence: 29021.60 (70%)	(Shape) Sphericity: 0.761 (-4.8%) (GLRLM) Gray level non-uniformity: 9.639 (7.6%) (GLCM) Cluster Prominence: 25621.04 (73%)
			
(Shape) Sphericity: 0.778 (GLRLM) Gray level non-uniformity: 8.714 (GLCM) Cluster Prominence: 83859.69	(Shape) Sphericity: 0.785 (0.9%) (GLRLM) Gray level non-uniformity: 10.099 (15%) (GLCM) Cluster Prominence: 69277.12 (-17%)	(Shape) Sphericity: 0.784 (0.8%) (GLRLM) Gray level non-uniformity: 13.287 (52%) (GLCM) Cluster Prominence: 57334.70 (-31%)	(Shape) Sphericity: 0.799 (2.7%) (GLRLM) Gray level non-uniformity: 9.698 (11%) (GLCM) Cluster Prominence: 23001.72 (-72%)

Fig. 1 Segmentation of three nodules from (left to right) 60-, 33-, 24-, and 15-mAs scans. Below each segmentation are listed high ICC shape (0.807), high ICC non-shape (0.966), and low ICC (0.207) feature values and increase in % compared to the 60-mAs feature value

considered as having poor reproducibility, values less than 0.75 as having moderate reproducibility, values between 0.75 and 0.9 as having good reproducibility, and values over 0.9 as having excellent reproducibility [17].

While the ICC metric is “ground truth agnostic,” treating every radiation dose level as being equivalent, it is arguably not the most optimal metric here. Due to the physical properties of computed tomography, a lower dose invariably leads to a worse signal-to-noise ratio. It is therefore likely that features extracted from lower-dose images contain the same or less information about the underlying biology of the nodule. We therefore performed an additional analysis where we treated the full-dose scan as a ground-truth observation. Features were scaled by the subtraction of the mean and the division by the standard deviation of the highest radiation dose (60 mAs) scans. To investigate how well ground-truth radiomic feature values can be obtained from lower-dose acquisitions using linear transformations, separate linear regression models were fitted for each feature and each reduced dose level. Feature values for 60 mAs were used as ground truth. These linear models were used to evaluate two metrics: *bias* and R^2 . Bias indicates the average deviation of feature values in a lower-dose setting from the average value in the full-dose (60 mAs) setting and is equal to the intercept term in a linear regression model. For each feature and for each dose level, the R^2 measures how much of the variation in ground-truth values can be explained using a linear correction of the lower-dose values. An R^2 value of 1 indicates that the values from the full-dose scan can be perfectly reconstructed from the lower-dose image using a linear model. A value of 0 indicates that it is impossible to reconstruct the ground-truth values from the lower-dose values using a linear model [18].

Results

Study population and radiomic feature extraction

Nineteen patients were included in the study, with ages ranging from 61 to 79 years (mean age: 67 years), of which 12 were male and 7 were female. Fifteen patients had lung nodules (35 in total) of which 3 were malignant. Of the fifteen patients, three patients (2 male and 1 female, with 3 nodules) were excluded because they presented with lung masses (diameter ≥ 3 cm) instead of lung nodules [19]. In total, 12 patients with 32 nodules with a median (IQR) diameter of 7.1 (6.1–9.6) mm were included for analysis in this study. In total, 1218 features were extracted from $32 \times 4 = 128$ nodules. A graphical abstract of three nodules with exemplary

feature values for the four different radiation doses is presented in Fig. 1.

Features considered stable (ICC)

Overall, only a minority of radiomic features were reproducible. From the 100 features without a filter applied, 15 features had excellent reproducibility ($ICC > 0.9$), 24 features had good reproducibility ($0.75 < ICC < 0.9$), 31 features had moderate reproducibility ($0.5 < ICC < 0.75$), and 30 features had poor reproducibility ($ICC < 0.5$). The top 30 ICC features are listed in Table 1. ICC values for all features are listed in Appendix 3. Of note, eight out of the top ten features with highest reproducibility were shape features. Overall, ten out of fourteen shape features were found to have an ICC value greater than 0.9 and can therefore be considered stable.

Table 1 Top 30/100 ICCs from original filter features

#	Feature	Value
1	original_glrlm_GrayLevelNonUniformity	0.966
2	shape_VoxelVolume	0.96
3	shape_MeshVolume	0.96
4	original_gldm_GrayLevelNonUniformity	0.952
5	shape_SurfaceArea	0.939
6	shape_MajorAxisLength	0.935
7	shape_LeastAxisLength	0.929
8	shape_Maximum2DDiameterSlice	0.928
9	shape_Maximum2DDiameterRow	0.92
10	shape_MinorAxisLength	0.913
11	shape_Maximum2DDiameterColumn	0.907
12	shape_Maximum3DDiameter	0.904
13	original_glcm_Imc2	0.902
14	original_glszm_GrayLevelNonUniformity	0.902
15	original_glcm_Imc1	0.901
16	original_glrlm_RunLengthNonUniformity	0.885
17	original_firstorder_90Percentile	0.881
18	original_firstorder_Median	0.856
19	original_glcm_Idn	0.855
20	original_glrlm_RunLengthNonUniformityNormalized	0.854
21	original_glrlm_RunPercentage	0.845
22	original_glszm_ZonePercentage	0.845
23	original_glrlm_ShortRunEmphasis	0.843
24	shape_SurfaceVolumeRatio	0.839
25	original_gldm_DependenceEntropy	0.836
26	original_gldm_DependenceNonUniformityNormalized	0.83
27	original_firstorder_Mean	0.83
28	original_glcm_Idmn	0.813
29	shape_Sphericity	0.807
30	original_glrlm_LongRunEmphasis	0.807

Effect of lower radiation dose on radiomic feature values (bias – R^2)

From the separate linear regression fits, bias and R^2 values were extracted. These values were plotted per filter category and per feature class. In general, features showed bias which increased with decreasing dose. In addition, for most features, R^2 values decreased for decreasing dose levels (Figs. 2 and 3). One percent of features had a negative slope fit. These features were omitted from the remainder of the analyses because this would imply that at a lower dose, the prognostic/diagnostic interpretation of a feature would be inverted, thus making these features unpractical

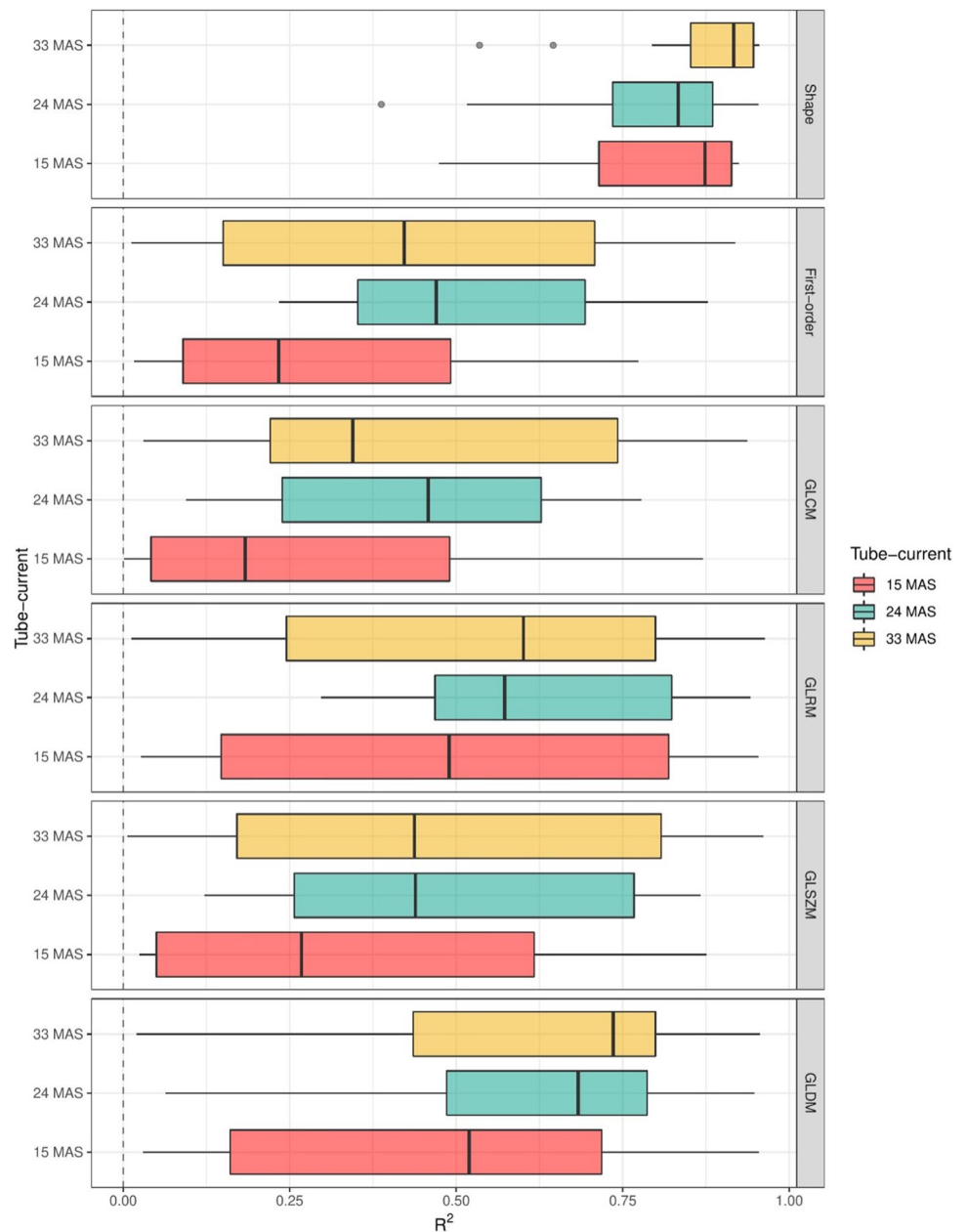
in a clinical setting. None of these features was from the subset of features without a filter applied. Negative slope features are listed in Appendix 4.

Bias increased and R^2 decreased with decreasing radiation dose (Figs. 2 and 3). In this analysis, the shape features were also found to have better correctability (higher R^2) compared to other features.

Robustness of features: radiomic feature classes (bias – R^2 , ICC)

To further analyze the robustness of radiomic features, the features were split in classes and bias versus R^2 was plotted as

Fig. 2 R^2 boxplot filter per feature class



a function of decreasing dose levels. The shape feature class was again found to be the most robust with the highest R^2 and the lowest bias (Fig. 4). An increasing trend in bias and a decreasing trend in R^2 were visible for all feature classes as a function of radiation dose. In other words, the difference from the mean of the high-dose (60 mAs) features was least for the shape feature class. Moreover, the error of shape features was fit best of all features by a linear model as a function of dose. All features were found to have an increasing difference from the mean of the high-dose features and a worse fit of the linear model, when dose level decreased.

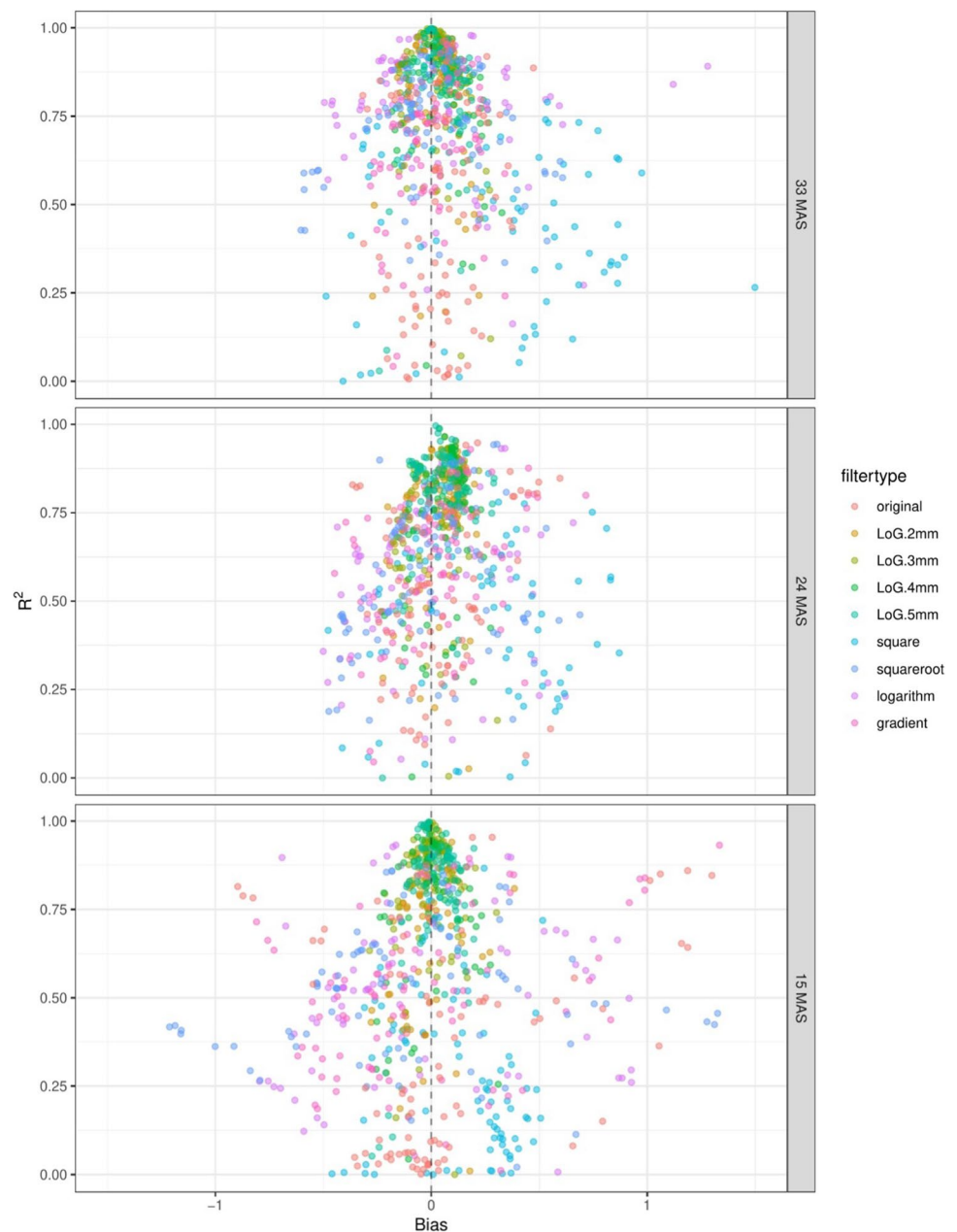
In addition, the ICC 3.1 was calculated [17]. ICC values per feature, split by feature class, are shown in Fig. 5.

Shape features had by far the highest ICC value of all feature classes, followed by GLRM features. This finding illustrates that shape features, followed by GLRM features, most strongly resemble each other in the different dose-level groups. Shape and first-order ICC, R^2 , and bias values are listed per feature in Tables 2 and 3.

Robustness of features: radiomic filters (bias – R^2)

Another possible variable that influences the reproducibility of radiomic features is the application of a filter to the image before feature extraction. Features calculated from filtered images were often less reproducible than those from

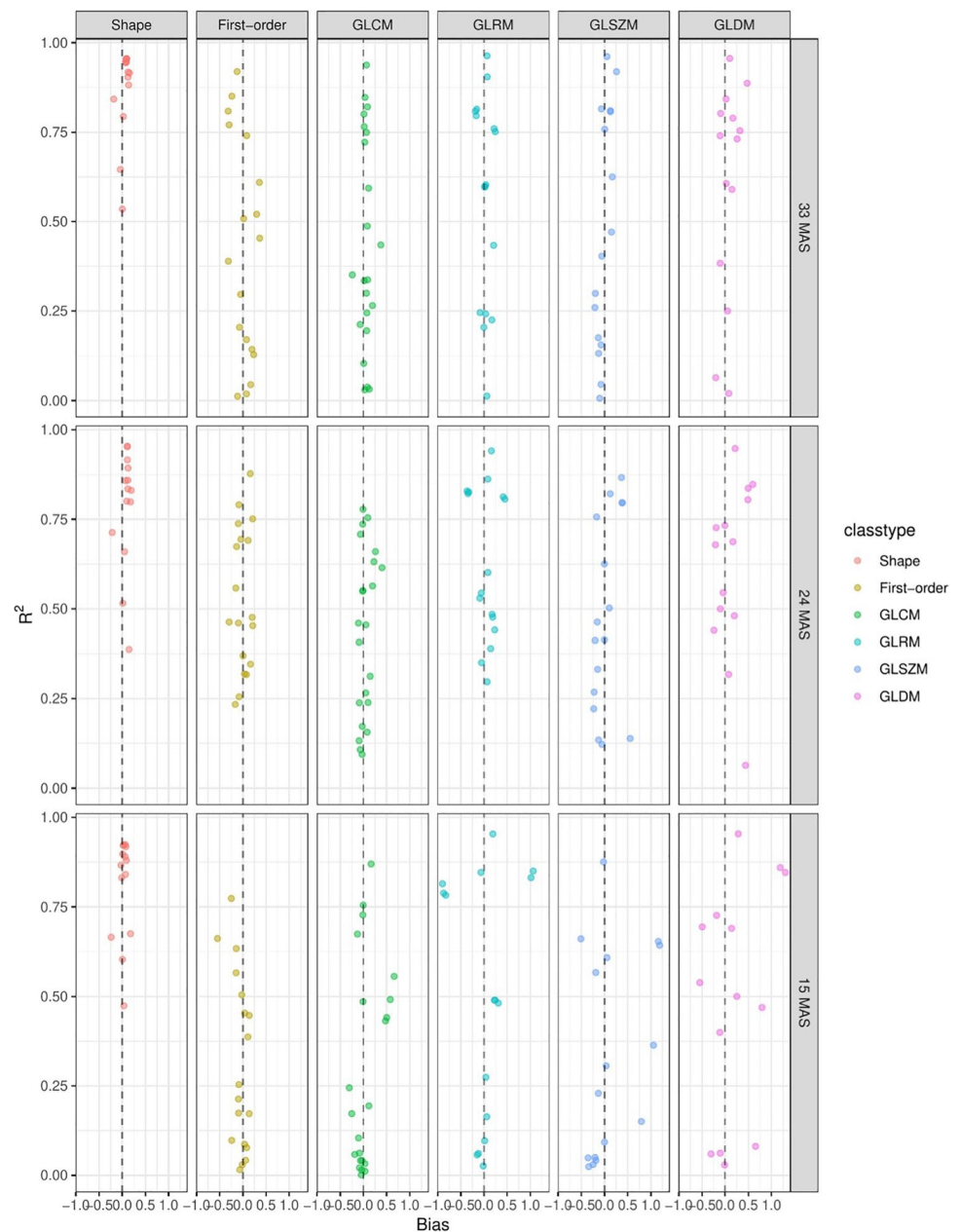
Fig. 3 Bias vs R^2 plotting for different milliampere-second levels looking at all features, colored by filters. High bias means that the value for this feature is on average higher than that for the reference dose of 60 mAs. High R^2 means that the deviation of feature values can be explained very well by a linear model taking into account only the dose (mAs)



the original image. This is demonstrated in Fig. 6, where R^2 and bias plots are shown for individual features, split by image filter. Figure 6 compares the original filter to the filter classes (Laplacian of Gaussian (LoG) (sigma 2, 3, 4, 5), square, square root, logarithm, and gradient).

The trend of decrease in R^2 and increase in bias were visible for all filters. Most filters were comparable to the original image regarding robustness of features. Wavelet, square, square root, logarithm, and gradient filters made the features less robust. The Laplacian of Gaussian filter seemed to make features remarkably more robust compared to the use of the original non-filtered images and other filters.

Fig. 4 Bias vs R^2 plotting for different milliampere-second levels and feature classes. High bias means that the value for this feature is on average higher than that for the reference dose of 60 mAs. R^2 is a statistic that will give some information about the goodness of fit of a model. High R^2 means that the deviation of feature values can be explained very well by a linear model taking into account only the dose (mAs)



Discussion

We performed an *in vivo*, intra-individual study on the robustness of radiomic features of pulmonary nodules as identified with computed tomography of the chest as a function of radiation dose levels. Except for shape features, we found that the majority of radiomic features are not stable against dose modulation. For a subset of features, it is possible to correct this variability by a simple linear model. However, the correction becomes increasingly less accurate at lower radiation doses.

Our finding that the majority of radiomic features are not stable against varying dose levels is concordant with

previously performed phantom studies that demonstrated a marked effect of CT tube current modulation on the value of several radiomic features [7, 8]. Our results are relevant for low-dose lung cancer screening. Globally, low-dose lung cancer screening is a growing trend and our findings underline the importance of standardizing the acquisition process. Ideally, screening and any follow-up examinations should be acquired on the same CT scanner with the same settings. Initiatives to standardize the process are being undertaken [1, 2].

The present results suggest that the most promising feature class regarding robustness is the shape feature class. Previous phantom studies have shown that shape features provide the most promising results regarding robustness against parameter variations (voxel geometry settings, dose

level, segmentation of ROI) [20, 21]. We found that first-order features were neither more robust nor more correctable by a linear model than other features. This is in contradiction to Hepp et al and Kim et al who found that first-order features were among the most stable in, respectively, a noise simulation study and a phantom study [10, 17].

From signal-processing theory, we know that a lower radiation dose introduces increasingly random noise to radiomic feature values. This is analogous to how the human visual system perceives lower quality. In other words, increased noise impairs the clinical value of radiomic features. For some features, a lower dose does not lead to noise but to systematic differences that are correctable. The error of a subset of unstable features can very well be explained by a simple

Fig. 5 ICC boxplot; high ICC values indicate robust features

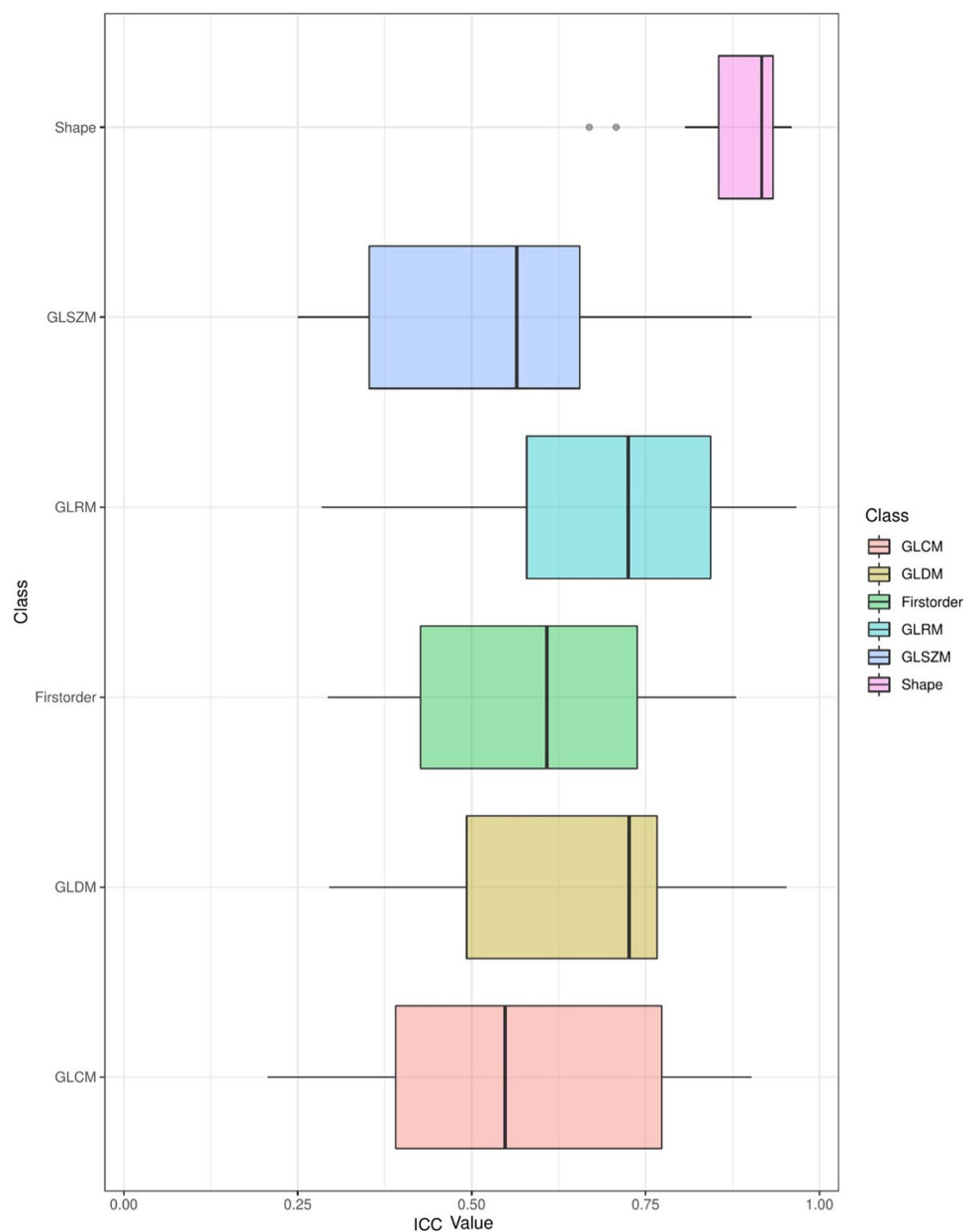


Table 2 ICC, R^2 , and bias for shape features

Feature	ICC	R^2	Bias	R^2	Bias	R^2	Bias
		15 mAs	15 mAs	24 mAs	24 mAs	33 mAs	33 mAs
VoxelVolume	0.96	0.92	0.02	0.95	0.11	0.96	0.09
Maximum3DDiameter	0.9	0.83	−0.01	0.84	0.12	0.94	0.08
MeshVolume	0.96	0.92	0.02	0.95	0.11	0.95	0.09
MajorAxisLength	0.93	0.92	0.08	0.86	0.12	0.95	0.07
Sphericity	0.81	0.68	0.18	0.66	0.05	0.79	0.02
LeastAxisLength	0.93	0.84	0.07	0.89	0.12	0.9	0.12
Elongation	0.67	0.6	0	0.39	0.14	0.53	0
SurfaceVolumeRatio	0.84	0.67	−0.24	0.71	−0.22	0.84	−0.19
Maximum2DDiameterSlice	0.93	0.92	0.06	0.83	0.19	0.92	0.16
Flatness	0.71	0.47	0.03	0.52	0.01	0.65	−0.04
SurfaceArea	0.94	0.9	0.01	0.92	0.11	0.95	0.08
MinorAxisLength	0.91	0.89	0.06	0.8	0.18	0.92	0.12
Maximum2DDiameterColumn	0.91	0.88	0.09	0.8	0.1	0.88	0.14
Maximum2DDiameterRow	0.92	0.87	−0.03	0.86	0.07	0.95	0.09

linear model (features with a high R^2). This is a promising result for more complicated correction methods. Zhovannik et al used an additive correction model to decrease error in 47 out of 62 feature values with at least a factor of 2 [7]. Wei et al used a 3D generative adversarial network to normalize reduced dose [22]. The decrease in error was significant for 8 out of 9 features. In addition, Mahon et al demonstrated the usage of the ComBat (combatting batch effect) harmonization algorithm, which greatly reduces the variation [20]. It remains to be seen if these methods can function as a uniform correcting method usable in clinical care. The vast

number of filters applied to the original image, apart from LoG, does not seem to generate more reproducible features. This raises the question whether there is any need for filters in the already vast amount of radiomic features extracted from the original image. Our finding that features derived from LoG-filtered images are more robust to dose variation is novel and warrants more investigation.

A unique advantage of this study is the radiographic imaging dataset. Fifteen patients underwent a CT scan at four different dose levels sequentially. The nature of the radiographic imaging dataset provides an opportunity to largely

Table 3 ICC, R^2 , and bias for first-order features

Feature	ICC	R^2	Bias	R^2	Bias	R^2	Bias
		15 mAs	15 mAs	24 mAs	24 mAs	33 mAs	33 mAs
InterquartileRange	0.43	0.48	−0.15	0.5	0	0.93	−0.05
Skewness	0.69	0.28	−0.05	0.24	−0.18	0.62	0.03
Uniformity	0.6	0.9	0.37	0.86	0.04	0.98	0.19
Median	0.86	0.51	0.45	0.82	0.23	0.62	0.1
Energy	0.7	0.67	−0.17	0.82	0.02	0.76	0.08
RobustMeanAbsoluteDeviation	0.43	0.6	−0.18	0.57	−0.04	0.91	−0.11
MeanAbsoluteDeviation	0.35	0.66	−0.31	0.57	−0.14	0.9	−0.15
TotalEnergy	0.68	0.66	−0.18	0.81	0	0.76	0.08
Maximum	0.76	0.44	−0.41	0.34	−0.42	0.78	−0.18
RootMeanSquared	0.75	0.7	−0.67	0.71	−0.44	0.9	−0.12
90Percentile	0.88	0.87	−0.14	0.56	−0.06	0.82	−0.1
Minimum	0.32	0.66	0.86	0.69	0.51	0.91	0.22
Entropy	0.54	0.88	−0.27	0.75	−0.06	0.97	−0.15
Range	0.39	0.44	−0.55	0.36	−0.5	0.79	−0.22
Variance	0.29	0.53	−0.41	0.47	−0.22	0.88	−0.18
10Percentile	0.61	0.67	0.75	0.7	0.44	0.88	0.16
Kurtosis	0.61	0.05	−0.08	0.05	−0.14	0.42	0.02
Mean	0.83	0.77	0.12	0.75	0.11	0.91	−0.01

isolate variables other than dose levels. To the best of our knowledge, this study is the first multi-dose *in vivo* study on lung nodule radiomic feature reproducibility.

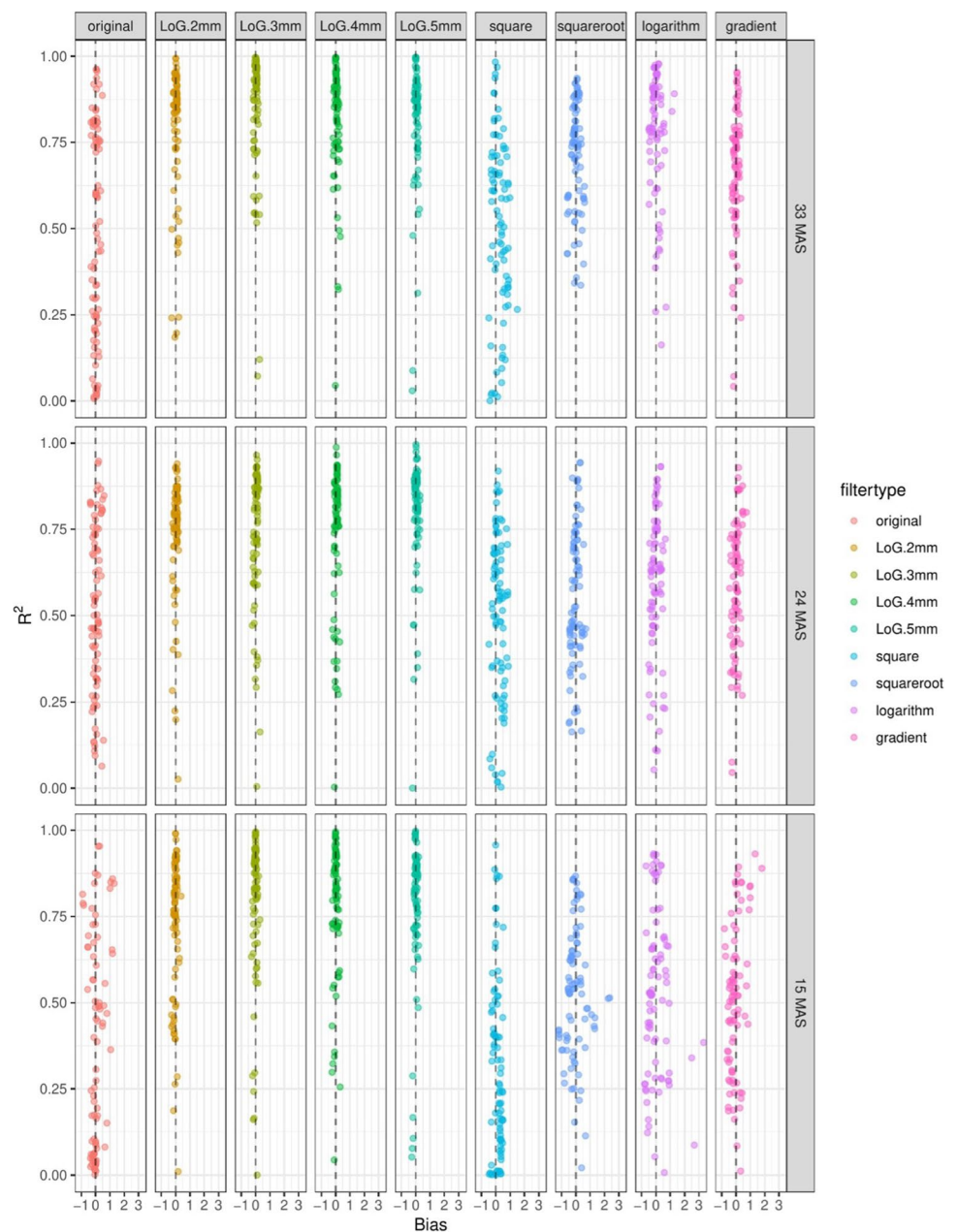
In general, we found shape features to be the most reproducible feature class. Yet, for a feature to be of clinical value, it must improve the diagnostic or prognostic value. Davey et al showed sphericity strongly correlates with overall survival of patients with lung cancer [21]. Yan et al showed that sphericity showed good ability in distinguishing adenocarcinoma from another lung cancer histological type using machine learning [23]. Liu et al found that a model for distinguishing benign from malignant lung nodules based on ten features, among which was the shape

feature sphericity, significantly outperformed a clinical variable-based model [23].

Shakir et al found that the shape feature surface volume ratio is most discriminative for nodule classification (benign vs malignant) out of 105 total features, using one-way ANOVA and three supervised selection algorithms [24]. Moreover, they found that the shape feature class had the highest relative contribution in nodule classification out of all the feature classes.

Yang et al selected seven features, among which were the shape features surface volume ratio and elongation, for the best diagnostic performance using hierarchical cluster analysis and the ReliefF method. The value of the conclusions on features with prognostic and/or diagnostic value

Fig. 6 Bias vs R^2 plotting for different milliampere-second levels and filters (original and wavelet). High bias means that the value for this feature is on average higher than that for the reference dose of 60 mAs. High R^2 means that the variation of feature values can be explained very well by a linear model taking into account only the dose (mAs)



is limited by slight differences in sets of radiomic features studied compared to this study. Future study needs to confirm if the radiomic features described in the current study have prognostic and/or diagnostic value.

This study has limitations. Manual delineation of the nodules was performed by only one investigator. Previous studies suggested that the standardization by using (semi-) automated segmentation methods provides more robust results [8, 25, 26]. However, the aim of the present study was to investigate if radiomics features are robust against dose modulation. We did not study whether features are sensitive to differences in ROI segmentation. Furthermore, it is known from the literature that this is indeed the case [25, 27, 28]. Therefore, we decided to have only one person segment all the scans. The extent to which segmentation differences interact with radiation dose reduction as to radiomics feature reproducibility is a very interesting question by itself and could very well be a direction for further research. Future studies should preferably be based on multiple delineations by multiple professionals or automation of segmentation. In addition, the high dimensionality of radiomic feature data hinders a simple presentation of results. To complicate the matter, a variety of presentation methods can be found in articles on the topic: ICC, concordance correlation coefficient (CCC), and coefficient of variation (COV) are all used interchangeably. This lack of consistency hinders comparison of results. For this study, we chose to plot bias and R^2 to intuitively visualize trends and calculate the ICC to quantify robustness. Our study counts a relatively small size (32) of nodules studied. This study did not investigate the prognostic or diagnostic value of radiomic features, only the stability of feature values over variations in radiation dose. We recommend further studies to investigate on the stability of radiomic features over different isolated variations such as manual delineation, bin width, or different reconstruction algorithms. The latter might be especially relevant as in a review by Reiazi et al radiation dose was found to be a disruptive parameter in all studies, whereas reconstruction algorithm appeared to be non-disruptive in about 50% of studies [11].

Also, we did not investigate the possible pre-processing of features or scans prior to feature calculation which might further enhance reproducibility [29]. Along the same vein, this study only investigated the reproducibility of radiomic features extracted from FBP constructed scans. Especially at lower milliampere-second levels, iterative reconstruction methods are used to decrease image noise. Shiri et al and Zhao et al showed that the variability and robustness of radiomic features in advanced reconstruction settings are feature-dependent [30, 31].

A solution to the possible lack of robustness of radiomic feature values is to standardize the process of feature extraction and possibly an (inter)national standardization of the clinical radiographic imaging setting. Although the latter seems a bridge too far currently, radiomic feature acquisition standardization initiatives are underway [2]. Finally, although the prespecified nature of radiomics features makes

them better explainable/connectable to the underlying biology, we cannot rule out that unsupervised deep learning techniques are less sensitive to variations in radiation dose.

In conclusion, a lower radiation dose introduces increasingly random noise and bias to radiomic feature values of pulmonary nodules. This noise can be corrected for by a linear model for a subset of features. We identified 15% of features as stable according to ICC, with shape as the most robust feature class.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00330-023-09643-8>.

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Tim Leiner.

Conflict of interest Martin J. Willeminck is a junior deputy editor of *European Radiology*, and Robbert W. van Hamersvelt is a member of the *European Radiology* Editorial Board. They have not taken part in the review or selection process of this article.

The other authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional review board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in <https://doi.org/10.1097/RCT.0000000000000408>.

Methodology

- Retrospective
- Observational
- Performed at one institution

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Amisha Malik P, Pathania M, Rathaur VK (2019) Overview of artificial intelligence in medicine. *J Family Med Prim Care* 8(7):2328–2331. https://doi.org/10.4103/jfmpc.jfmpc_440_19

2. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295(2):328–338. <https://doi.org/10.1148/radiol.2020191145>
3. Fitzmaurice C, Allen C, Barber RM et al (2017) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the Global Burden of Disease Study. *JAMA Oncol* 3(4):524–548. <https://doi.org/10.1001/jamaoncol.2016.5688>
4. Kubo T, Ohno Y, Kauczor HU, Irich, Hatabu H (2014) Radiation dose reduction in chest CT—review of available options. *Eur J Radiol* 83(10):1953–1961. <https://doi.org/10.1016/j.ejrad.2014.06.033>
5. Solomon JB, Li X, Samei E (2013) Relating noise to image quality indicators in CT examinations with tube current modulation. *AJR Am J Roentgenol* 200(3):592–600
6. Larue RTHM, van Timmeren JE, de Jong EEC et al (2017) Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol* 56(11):1544–1553. <https://doi.org/10.1080/0284186X.2017.1351624>
7. Zhovannik I, Bussink J, Traverso A et al (2019) Learning from scanners: bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol* 19:33–38. <https://doi.org/10.1016/j.ctro.2019.07.003>
8. Hepp T, Othman A, Liebgott A, Kim JH, Pfannenber C, Gatidis S (2020) Effects of simulated dose variation on contrast-enhanced CT-based radiomic analysis for non-small cell lung cancer. *Eur J Radiol* 124:108804. <https://doi.org/10.1016/j.ejrad.2019.108804>
9. Meyer M, Ronald J, Vernuccio F et al (2019) Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. *Radiology* 293(3):583–591
10. Lu L, Sun SH, Afran A et al (2021) Identifying robust radiomics features for lung cancer by using in-vivo and phantom lung lesions. *Tomography* 7(1):55–64
11. Reiazi R, Abbas E, Famiyeh P et al (2021) The impact of the variation of imaging parameters on the robustness of computed tomography radiomic features: a review. *Comput Biol Med* 133:104400. <https://doi.org/10.1016/j.compbiomed.2021.104400>
12. den Harder AM, Willeminck MJ, van Hamersvelt RW, et al (2016) Pulmonary nodule volumetry at different low computed tomography radiation dose levels with hybrid and model-based iterative reconstruction: a within patient analysis. *J Comput Assist Tomogr* 40(4). https://journals.lww.com/jcat/Fulltext/2016/07000/Pulmonary_Nodule_Volumetry_at_Different_Low.14.aspx. Accessed Jan 2021
13. den Harder AM, Willeminck MJ, van Hamersvelt RW et al (2016) Effect of radiation dose reduction and iterative reconstruction on computer-aided detection of pulmonary nodules: intra-individual comparison. *Eur J Radiol* 85(2):346–351. <https://doi.org/10.1016/j.ejrad.2015.12.003>
14. Zhu L, Kolesov I, Gao Y, Kikinis R, Tannenbaum A (2014) An effective interactive medical image segmentation method using fast GrowCut. In: *Int Conf Med Image Comput Comput Assist Interv. Workshop Interact Meth Vol 17*
15. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77(21):e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
16. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86(2):420–428. <https://doi.org/10.1037//0033-2909.86.2.420>
17. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15(2):155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
18. Hamilton DF, Ghert M, Simpson AHRW (2015) Interpreting regression models in clinical outcome studies. *Bone Joint Res* 4(9):152–153. <https://doi.org/10.1302/2046-3758.49.2000571>
19. Nair A, Devaraj A, Callister MEJ, Baldwin DR (2018) The Fleischner Society 2017 and British Thoracic Society 2015 guidelines for managing pulmonary nodules: keep calm and carry on. *Thorax* 73(9):806. <https://doi.org/10.1136/thoraxjnl-2018-211764>
20. Mahon RN, Ghita M, Hugo GD, Weiss E (2020) ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Phys Med Biol* 65(1):15010. <https://doi.org/10.1088/1361-6560/ab6177>
21. Davey A, van Herk M, Faivre-Finn C, Mistry H, McWilliam A (2020) Is tumour sphericity an important prognostic factor in patients with lung cancer? *Radiother Oncol* 143:73–80. <https://doi.org/10.1016/j.radonc.2019.08.003>
22. Wei L, Lin Y, Hsu W (2020) Using a generative adversarial network for CT normalization and its impact on radiomic features. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 844–848. <https://doi.org/10.1109/ISBI45749.2020.9098724>
23. Yan M, Wang W (2020) A non-invasive method to diagnose lung adenocarcinoma. *Front Oncol* 10:602. <https://doi.org/10.3389/fonc.2020.00602>
24. Shakir H, Rasheed H, Khan TMR et al (2020) Radiomic feature selection for lung cancer classifiers. *Journal of Intelligent and Fuzzy Systems*. 38:5847–5855. <https://doi.org/10.48550/arXiv.2003.07098>
25. Haarbuerger C, Müller-Franzes G, Weninger L, Kuhl C, Truhn D, Merhof D (2020) Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. *Sci Rep* 10(1):12688. <https://doi.org/10.1038/s41598-020-69534-6>
26. Haarbuerger C, Schock J, Truhn D, et al (2020) Radiomic feature stability analysis based on probabilistic segmentations. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. :1188–1192. <https://doi.org/10.1109/ISBI45749.2020.9098674>
27. Pavic M, Bogowicz M, Würms X et al (2018) Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol* 57(8):1070–1074. <https://doi.org/10.1080/0284186X.2018.1445283>
28. Kalpathy-Cramer J, Mamomov A, Zhao B et al (2016) Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography* 2(4):430–437. <https://doi.org/10.18383/j.tom.2016.00235>
29. Bologna M, Corino VDA, Montin E et al (2018) Assessment of stability and discrimination capacity of radiomic features on apparent diffusion coefficient images. *J Digit Imaging* 31(6):879–894. <https://doi.org/10.1007/s10278-018-0092-9>
30. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A (2017) The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol* 27(11):4498–4509. <https://doi.org/10.1007/s00330-017-4859-z>
31. Zhao W, Zhang W, Sun Y et al (2019) Convolution kernel and iterative reconstruction affect the diagnostic performance of radiomics and deep learning in lung adenocarcinoma pathological subtypes. *Thorac Cancer* 10(10):1893–1903. <https://doi.org/10.1111/1759-7714.13161>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.