


ORIGINAL ARTICLE

Open Access



Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters

Lisa Rinaldi^{1,2}, Simone P. De Angelis³, Sara Raimondi^{3*} , Stefania Rizzo^{4,5}, Cristiana Fanciullo⁶, Cristiano Rampinelli⁷, Manuel Mariani², Alessandro Lascialfari², Marta Cremonesi¹, Roberto Orecchia⁸, Daniela Origgi^{9†} and Francesca Botta^{9†}

Abstract

Background: We investigated to what extent tube voltage, scanner model, and reconstruction algorithm affect radiomic feature reproducibility in a single-institution retrospective database of computed tomography images of non-small-cell lung cancer patients.

Methods: This study was approved by the Institutional Review Board (UID 2412). Images of 103 patients were considered, being acquired on either among two scanners, at 100 or 120 kVp. For each patient, images were reconstructed with six iterative blending levels, and 1414 features were extracted from each reconstruction. At univariate analysis, Wilcoxon-Mann-Whitney test was applied to evaluate feature differences within scanners and voltages, whereas the impact of the reconstruction was established with the overall concordance correlation coefficient (OCCC). A multivariable mixed model was also applied to investigate the independent contribution of each acquisition/reconstruction parameter. Univariate and multivariable analyses were combined to analyse feature behaviour.

Results: Scanner model and voltage did not affect features significantly. The reconstruction blending level showed a significant impact at both univariate analysis (154/1414 features yielding an OCCC < 0.85) and multivariable analysis, with most features (1042/1414) revealing a systematic trend with the blending level (multiple comparisons adjusted $p < 0.05$). Reproducibility increased in association to image processing with smooth filters, nonetheless specific investigation in relation to clinical endpoints should be performed to ensure that textural information is not removed.

Conclusions: Combining univariate and multivariable models is allowed to identify features for which corrections may be applied to reduce the trend with the algorithm and increase reproducibility. Subsequent clustering may be applied to eliminate residual redundancy.

Keywords: Carcinoma (non-small-cell lung), Image processing (computer-assisted), Machine learning, Reproducibility of results, Tomography (x-ray computed)

* Correspondence: sara.raimondi@ieo.it

†Daniela Origgi and Francesca Botta are co-last authors and both share co-senior authorship.

³Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy

Full list of author information is available at the end of the article



© The Author(s) under exclusive licence to European Society of Radiology. 2022 **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/>.

Key points

- Scanner and voltage did not affect features significantly.
- Reconstruction algorithm blending levels impact on the reproducibility of features.
- Combination of multiple analyses may help to characterise feature behaviour.

Background

Current clinical practice relies on radiological imaging for the diagnosis, evaluation, and monitoring of diseases. Radiomics is an emerging discipline which aims to add further quantitative objectivity to the visual interpretation of the medical images performed by the physicians [1, 2]. A huge variety of mathematical descriptors, named *radiomic features*, can be calculated from images quantifying different aspects of the tumour shape and texture. Radiomics investigates the ability of some of such features to characterise clinical properties of the lesions [3]. Under the assumption that the features might capture relevant information not discriminated by the human eye [4], radiomics attempts to become a precious tool to support personalised clinical decisions [5].

Importantly, radiomics can be applied successfully in the clinical practice only if the radiomic-based predictive models are robust and generalisable. To this aim, radiomic features must not be biased by any variation in the image signal except for that ascribable to actual biological processes [6, 7]. Conversely, it was extensively observed that the use of different parameters during acquisition and post-acquisition may modify the image signal with a significant impact on the value of the radiomic features, even when the diagnostic quality of the image is maintained [8–11].

Quite often image databases collected for radiomic research are highly heterogeneous, including images obtained with different scanners, acquisition protocols and post-acquisition techniques [1, 12, 13]. Among others, a recent study from our group [14] confirmed the possible confounding factor of reconstruction algorithms. In this study, we built a predictive model, identifying a possible association of the radiomic and clinical information with the lymph node status and the overall survival in 270 patients with lung cancer. Through an analysis of the variance on 422 selected radiomic features, we found that 254 of them differed significantly between the two types of reconstruction algorithm: iterative reconstruction (IR) and filtered backprojection (FBP). In these cases, it is fundamental to take the reproducibility issue into account, by either disregarding or properly correcting and harmonising the features significantly affected by the different imaging procedures [15–17].

Focusing on computed tomography (CT), previous studies investigated the radiomic features variability related to different acquisition parameters (tube current [18], slice thickness [19–22] and tube voltage peak [18]), reconstruction techniques [19, 20, 22–26], segmentation of the volume of interest [27–29], and post-processing techniques [30–33]. Quite often, however, the list of reproducible features obtained in one study is not fully provided or is not directly exportable to a different database of images, if obtained with different equipment or image settings.

In this paper, we faced the reproducibility issue in a retrospective database of CT images available at our Institute for patients affected by non-small-cell lung cancer (NSCLC). We aimed to contribute to the existing literature first by identifying the list of reproducible features for radiomic analysis of NSCLC CT studies, and, most importantly, by suggesting an integration of different metrics, which can improve the interpretation of reproducibility results and can be replicated in other scenarios.

Methods

Patients and ethics issues

Patients undergoing diagnostic chest contrast-enhanced CT imaging at our Institute between January 2019 and December 2019 were retrospectively selected. Inclusion criteria were the availability of CT raw-data, CT being acquired with the institutional standard protocol and with beam energy set to either 100 or 120 kVp, and histologically proven diagnosis of NSCLC. Exclusion criteria were tumour volume smaller than 5 cm³ or larger than 200 cm³. The Institutional Review Board approved the study (UID 2412) waiving the need for informed consent.

Image acquisition and reconstruction

Contrast-enhanced CT images were acquired using either Discovery CT750 HD or Optima CT660 scanner (General Electric Healthcare, Wisconsin, USA) according to the current institutional standard protocol (acquisition: helical acquisition, 2.5 mm slice thickness and spacing, automatic tube current modulation, tube voltage set to 100, 120 or 140 kVp according to patient body mass index, noise index suitably optimised for each voltage to provide comparable image quality; reconstruction with standard convolution kernel, adaptive statistical iterative reconstruction (ASIR) algorithm with 60% blending level on Discovery CT750 HD and 50% blending level on Optima CT660). At our institute, two types of iodinated-contrast medium are usually injected, Visipaque® 320 (General Electric Healthcare, Wisconsin, USA) or Ultravist® 370 (Bayer Healthcare, Leverkusen, Germany), and the volume of the contrast medium is

selected depending on contrast medium concentration and patient weight.

Previously, during the optimisation process, different blending levels were tested and appear in our retrospective databases. To replicate and investigate this variability, the CT images (portal phase series) of the patients included in this study were reconstructed applying each time a different IR blending level: 0% (equivalent to FBP), 20%, 40%, 50%, 60% and 80% (referred to as IR20, IR40, IR50, IR60 and IR80, respectively).

Tumour segmentation

One pulmonary lesion for each patient was contoured manually slice by slice on the series used for clinical reporting (AWSer 3.2 Ext. 2.0 tool, General Electric Healthcare, Wisconsin, USA). The so obtained volume of interest was used for the radiomic analysis of all the six reconstructions, inherently co-registered. This allowed us to investigate the impact of reconstruction algorithms avoiding possible biases that might have occurred if the segmentation was repeated separately on each reconstructed image.

Tumours were contoured by three operators with similar experience (more than 7 years of experience) after agreement on segmentation criteria and settings. Radiologists trained among each other to reach a consensus on the segmentation procedure, including window setting for visualisation (1500 Hounsfield units, HU, width and - 600 HU level for lung window, 350 HU width and 40 HU level for mediastinal window, depending on lesion localisation), exclusion of the vessels, and inclusion of opacity on the lesion edge.

Radiomic feature extraction

The radiomic features were computed through the open-source package Pyradiomics v. 2.2.0 [34], from each of the six reconstructed images for each patient. Radiomic features were extracted considering the following seven categories: shape; first order; grey level co-occurrence matrix; grey level run length matrix; grey level size zone matrix; neighbouring grey tone difference matrix; and grey level dependence matrix.

According to IBSI recommendations [35], before feature computation image resampling in the axial plane (Pyradiomics B-Spline interpolator, 'sitkBSpline' [36, 37]) and voxel intensity discretisation (25 HU fixed bin [7, 34, 38, 39]) were applied.

Shape, intensity (*first order*) and *texture* features were calculated, both from original images without filtering and after applying the wavelet filter (order 1 Coiflet, Pyradiomics default [1, 7, 34, 40]) and Laplacian of Gaussian (LoG) filter with different values of Gaussian standard deviation (*sigma*: 0.5, 1.0, 1.5, 2.5 and 5.0 mm [20, 34, 41–43]). The names of the features will be presented with the suffix "original", "Wavelet" or "LoG",

followed by the feature category and the feature name. Additional details on extracted feature categories and parameters set for calculation are reported in [Supplementary Methods](#).

Features from the *shape* category will be included only in the "original" group since they are identical for original and filtered images.

Statistical analysis

Clinical similarity between patient and tumour characteristics (age, volume, gender, side, position, tumour type, previous therapy and pTNM stage), according to scanner and tube voltage, was evaluated with χ^2 or Fisher exact test for categorical variables, and with Wilcoxon-Mann-Whitney test for continuous variables.

Univariate analysis to evaluate differences in feature values within CT scanners (Optima CT660 versus Discovery CT750 HD) and within tube voltages (100 versus 120 kVp) was performed with Wilcoxon-Mann-Whitney test using feature values obtained from FBP and IR60 images (this latter chosen as representative, being in the middle of the IR blending level interval investigated).

The overall concordance between the six different settings for the reconstruction algorithm on the same patient was evaluated for each feature with the overall concordance correlation coefficient (OCCC) [44]. An OCCC threshold equal to 0.85 was used to classify features affected (OCCC < 0.85) or not (OCCC \geq 0.85) by the IR blending level applied during reconstruction [20, 45, 46].

Additionally, to investigate the independent contribution of each acquisition and reconstruction parameter on feature variation, a multivariable mixed model was used, including subjects as random effect to take into account within-subject variation for the six reconstructions, and adjusting by clinical volume. For this analysis, FBP was set as reference category, and one model coefficient and *p* value for each IR blending level was calculated for comparison with FBP. All *p* values were corrected with the false discovery rate (FDR) method [47] to properly account for multiple testing; adjusted *p* values < 0.05 were considered statistically significant.

For a deeper understanding of feature dependence on reconstruction algorithm settings, features were classified in four groups.

Group 1, with OCCC \geq 0.85 and mixed model FDR-adjusted *p* value < 0.05. Over-threshold OCCC indicates that feature variations among the different reconstruction settings for each patient are small in comparison to the variations observed in the entire dataset (differences among patients). The significant *p* value of the multivariable mixed model indicates that such small variations follow systematically the same trend for almost all the patients in the dataset. Hence, features belonging to this category change slightly when

modifying the reconstruction setting, and the trend of such variation can be predicted and corrected, since it is similar for almost all patients.

Group 2, with $OCCC \geq 0.85$ and mixed model FDR-adjusted p value ≥ 0.05 . When changing the reconstruction setting, the features vary slightly (or do not vary at all if $OCCC = 1$) in comparison to the whole dataset variations, but the non-significant p value of the mixed model indicates that the trend of such variation (if any) is not systematic among patients, but random.

Group 3, with $OCCC < 0.85$ and mixed model FDR-adjusted p value < 0.05 . The low $OCCC$ value indicates that the feature variation when changing the reconstruction setting is not negligible in comparison to the variations observed in the whole dataset. The trend of such variations is systematically the same for almost all patients.

Group 4, with $OCCC < 0.85$ and mixed model FDR-adjusted p value ≥ 0.05 . The features exhibit a relevant variation in comparison to the variations observed in the whole dataset, but the sign and entity of such variations change randomly among patients.

As sensitivity analysis, pair differences were calculated with Wilcoxon signed rank-test to compare algorithms among them (not necessarily with FBP as reference).

The whole analysis was also performed on a subgroup of IR blending levels (IR40, IR50, IR60 and IR80), taking the IR40 as a reference, in order to provide results also in a setting more representative of the current clinical applications.

Finally, the features extracted from original images were clustered according to a minimum intra-cluster correlation criterion (Spearman's $|\rho| \geq 0.75$) to quantify feature redundancy.

All analyses were performed with R (v. 4.0.0) [48], and tests were two-sided.

Results

Among 163 patients selected for the availability of CT raw-data, 103 (59 men, mean age 71 years; 44 women, mean age 67 years) fulfilled the remaining enrolment criteria and were included in the study: 50 (49%) imaged on Optima CT660 scanner (50% at 100 kVp, 50% at 120 kVp); 53 (51%) imaged on Discovery CT750 HD scanner (51% at 100 kVp, 49% at 120 kVp), resulting in four populations according to scanner and tube voltage (Figure S1).

The baseline clinical characteristics are summarised in Table 1, as long as the p values for the comparison of clinical characteristics between the two scanners and the two tube voltage patient populations. No statistically significant difference was found, confirming clinical similarity of the four populations. For the subgroup of patients with available information on pTNM stage and

grading, no statistically significant difference was observed among the populations (results not shown).

A total of 1414 radiomic features were extracted, including 154 from original images (14 *shape*, 17 *first order* and 123 *texture* features), 560 from wavelet-filtered images (68 *first order* and 492 *texture*) and 700 from LoG-filtered images (85 *first order* and 615 *texture*). The full feature list is reported in Table S1 along with the 33 groups in which the original image features were clustered.

Scanner model and tube voltage

Forty-four features were significantly different according to scanner and/or tube voltage, either at univariate or multivariable (mixed model) analysis or both (Table S2). Focusing on multivariable analysis, only 5 features (1 from *shape* category and 4 from *texture* category and wavelet-filtered images) showed significant dependence on tube voltage, and 1 (*shape_SurfaceArea*) on scanner (Table 2).

Reconstruction algorithm

In order to evidence the impact of the reconstruction algorithm on the image texture, we reported an example of two reconstructions (FBP and IR80) of the same lesion (Fig. 1).

From the concordance analysis between the different reconstruction settings, we obtained that 16/154 features (10%) had small reproducibility ($OCCC < 0.85$), all in *texture* categories, in case of features from original images (Fig. 2a). Features from *grey level run length matrix* category were mostly affected by reconstruction algorithm setting. In case of wavelet-filtered images (Fig. 2b), 116/560 features (21%) yielded $OCCC < 0.85$, mostly (51%) from the HH-wavelet group, whereas LL-wavelet features exhibited the highest concordance. Features from LoG-filtered images showed the highest reproducibility in all feature categories and for each value of sigma parameter (Fig. 2c): only 22/700 features (3%) yielded $OCCC < 0.85$. The analogous results obtained when restricting the analysis to the IR40-IR80 range are reported in Figure S2, with 6/154 (4%), 22/560 (4%) and 6/700 (1%) features yielding $OCCC < 0.85$ for the original, wavelet and LoG features, respectively. Table 3 reports the median $OCCC$ for each image type and feature category, both for the main analysis and, in parentheses, for the subanalysis restricted to the IR40-IR80 range. Full results ($OCCC$ value obtained for each feature) are reported in Tables S3, S4 and S5 along with the FDR-adjusted p values obtained from the multivariable mixed model.

According to the multivariable mixed model, 110/140 (78.5%) features from original images (*shape* features excluded), 462/560 (82.5%) from wavelet-filtered images and

Table 1 Baseline characteristics of the study population.

Variables	Overall cohort (n = 103)	Scanner Optima CT660 (n = 50)	Scanner Discovery CT750 HD (n = 53)	p value (scanner)	Tube voltage 120 kVp (n = 51)	Tube voltage 100 kVp (n = 52)	p value (kVp)
Gender							
Male	59 (57%)	28 (56%)	31 (58%)	0.798 ^a	33 (65%)	26 (50%)	0.131 ^a
Female	44 (43%)	22 (44%)	22 (42%)		18 (35%)	26 (50%)	
Age							
Mean (median)	69.2 (70)	69.4 (70)	68.9 (69)	0.498 ^c	69.8 (70)	68.6 (68.5)	0.251 ^c
IQR	(64–75)	(65–75.3)	(62–74.5)		(64–76)	(62–74.8)	
Side							
Right	60 (58%)	31 (62%)	29 (55%)	0.454 ^a	31 (61%)	29 (56%)	0.606 ^a
Left	43 (42%)	19 (38%)	24 (45%)		20 (39%)	23 (44%)	
Position							
Upper	63 (64%)	33 (69%)	30 (60%)	0.360 ^b	30 (61%)	33 (67%)	0.731 ^b
Medium	1 (1%)	1 (2%)	0 (0%)		1 (2%)	0 (0%)	
Lower	29 (30%)	13 (27%)	16 (32%)		16 (33%)	13 (27%)	
Mixed	5 (5%)	1 (2%)	4 (8%)		2 (4%)	3 (6%)	
Volume (cm³)							
Mean (median)	46.4 (39.1)	44.2 (40.6)	48.5 (38.1)	0.843 ^c	52.1 (42)	40.9 (36.7)	0.181 ^c
IQR	(19.1–62.8)	(19–54.7)	(19.5–71.9)		(20.7–67.9)	(18.4–56.2)	
Histological type							
Adenocarcinoma	83 (82%)	38 (78%)	45 (87%)	0.580 ^b	40 (78%)	43 (86%)	0.380 ^b
Squamous cell carcinoma	16 (16%)	10 (20%)	6 (11%)		9 (18%)	7 (14%)	
Neuroendocrine	2 (2%)	1 (2%)	1 (2%)		2 (4%)	0 (0%)	
Previous therapy							
No	75 (74%)	38 (76%)	37 (73%)	0.692 ^a	33 (66%)	42 (82%)	0.060 ^a
Yes	26 (26%)	12 (24%)	14 (27%)		17 (34%)	9 (18%)	
Scanner							
Optima CT660	50 (49%)	–	–	–	25 (49%)	25 (48%)	0.924 ^a
Discovery CT750 HD	53 (51%)	–	–	–	26 (51%)	27 (52%)	
Tube voltage (kVp)							
120	51 (50%)	25 (50%)	26 (49%)	0.924 ^a	–	–	–
100	52 (50%)	25 (50%)	27 (51%)		–	–	

^a χ^2 test

^bFisher's exact test

^cWilcoxon-Mann-Whitney test. Missing data: histological type (n = 2); previous therapy (n = 2); position (n = 5). IQR Interquartile range

Table 2 FDR-adjusted p values for univariate and multivariable analysis for the effect of scanner and tube voltage

Features	Scanner (univar) FBP	Scanner (univar) IR60	Tube voltage (univar) FBP	Tube voltage (univar) IR60	Scanner (mixed)	Tube voltage (mixed)
shape_SurfaceArea	0.897	0.960	0.735	0.695	0.027	0.886
shape_VoxelVolume	0.936	0.960	0.784	0.695	0.190 [°]	< 0.001 [°]
Wavelet-glszm_SizeZoneNonUniformityNormalized*	0.264	0.905	0.005	0.016	0.996	0.005
Wavelet-glszm_SmallAreaEmphasis*	0.264	0.905	0.006	0.016	0.996	0.005
Wavelet-glcm1_Correlation*	0.462	0.905	0.144	0.130	0.561	0.018
Wavelet-glcm1_InverseVariance*	0.231	0.905	0.004	0.097	0.309	0.012

Only the features with significant FDR-adjusted p values at multivariate analysis

[°]HH filter

[°]In the model with VoxelVolume as the dependent variable, clinical volume was not used as independent predictor. FBP filtered backprojection, FDR false discovery rate, IR iterative reconstruction

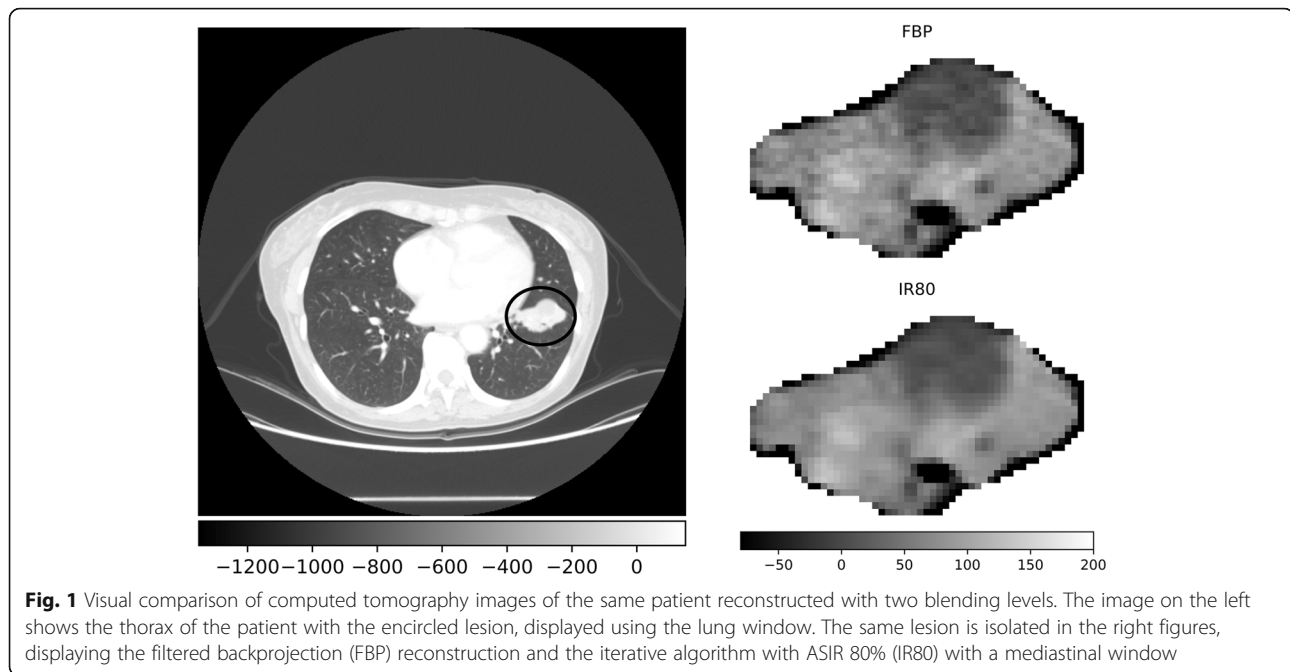


Fig. 1 Visual comparison of computed tomography images of the same patient reconstructed with two blending levels. The image on the left shows the thorax of the patient with the encircled lesion, displayed using the lung window. The same lesion is isolated in the right figures, displaying the filtered backprojection (FBP) reconstruction and the iterative algorithm with ASIR 80% (IR80) with a mediastinal window

470/700 (67%) from LoG-filtered images were significantly affected by IR setting (mixed model FDR-adjusted p value < 0.05). Similar results were obtained for the subanalysis restricted to the IR40-IR80 range: 112/140 (80%) for original images, 457/560 (82%) for wavelet-filtered images and 421/700 (60%) for LoG-filtered images.

We combined the results of the two metrics adopted for the reproducibility analysis and divided the features in four groups, as described in the “Methods” section (“Statistical analysis” section). One representative feature for each group was selected and displayed in Fig. 3 to highlight the different behaviours of the features falling in the different groups. To this aim, we plotted the absolute value of these features when increasing the reconstruction blending levels for the four patient populations, each line representing a different patient.

We found that the majority of the features fall in group 1 (OCCC ≥ 0.85 and p value < 0.05), suggesting the capability of the features to capture the gradual smoothing effect of the increasing IR strength on the image texture, with a similar trend for all the patients. In contrast, group 4 is the less populated. In Fig. 4, we reported some examples of these findings, by plotting the OCCC value versus the mixed model FDR-adjusted p value (each point in the graph representing a feature) for six cases: original (Fig. 4a), wavelet-filtered (Fig. 4b) and LoG-filtered (Fig. 4c) images, in each case including the two extreme configurations of IR blending level (IR20 and IR80) versus FBP. The red lines divide each plot in four quadrants, corresponding to the four groups described in the “Methods” section (“Statistical analysis” section). The percentage of features falling in each group

for the six cases is reported in Table 4, whereas Table S6 reports the corresponding results for the subanalysis (IR50 and IR80 versus IR40).

The pair comparisons among reconstruction settings are reported in Figure S3 for the features obtained from original images. The number of features with significantly different values between two algorithms ranges from 117 (84%, excluding *shape* category) to 128 (91%), with the number of poorly reproducible features increasing when increasing the IR blending level interval (considering FBP as more similar to IR20).

Discussion

The main findings of this study are related to the influence of reconstruction setting on the value of radiomic features, and its interpretation. Our findings in relation to the dependence on scanner and tube voltage (not statistically significant in our sample) basically confirm previous results [18, 49–51].

Besides confirming that the IR blending level has a significant impact on the value of a set of features extracted from CT images of patients affected by NSCLC [20, 23, 24, 52], we provided feature-by-feature results which might be conveniently compared with similar findings obtained on different dataset (images of clinically comparable cases obtained at different Institutes, with different scanner models, acquisition and reconstruction settings) to verify if the subset of reproducible radiomic features is coherent among different samples.

In addition, we introduced a novel approach to investigate and handle the dependency of each feature value on the reconstruction setting. By joining two different

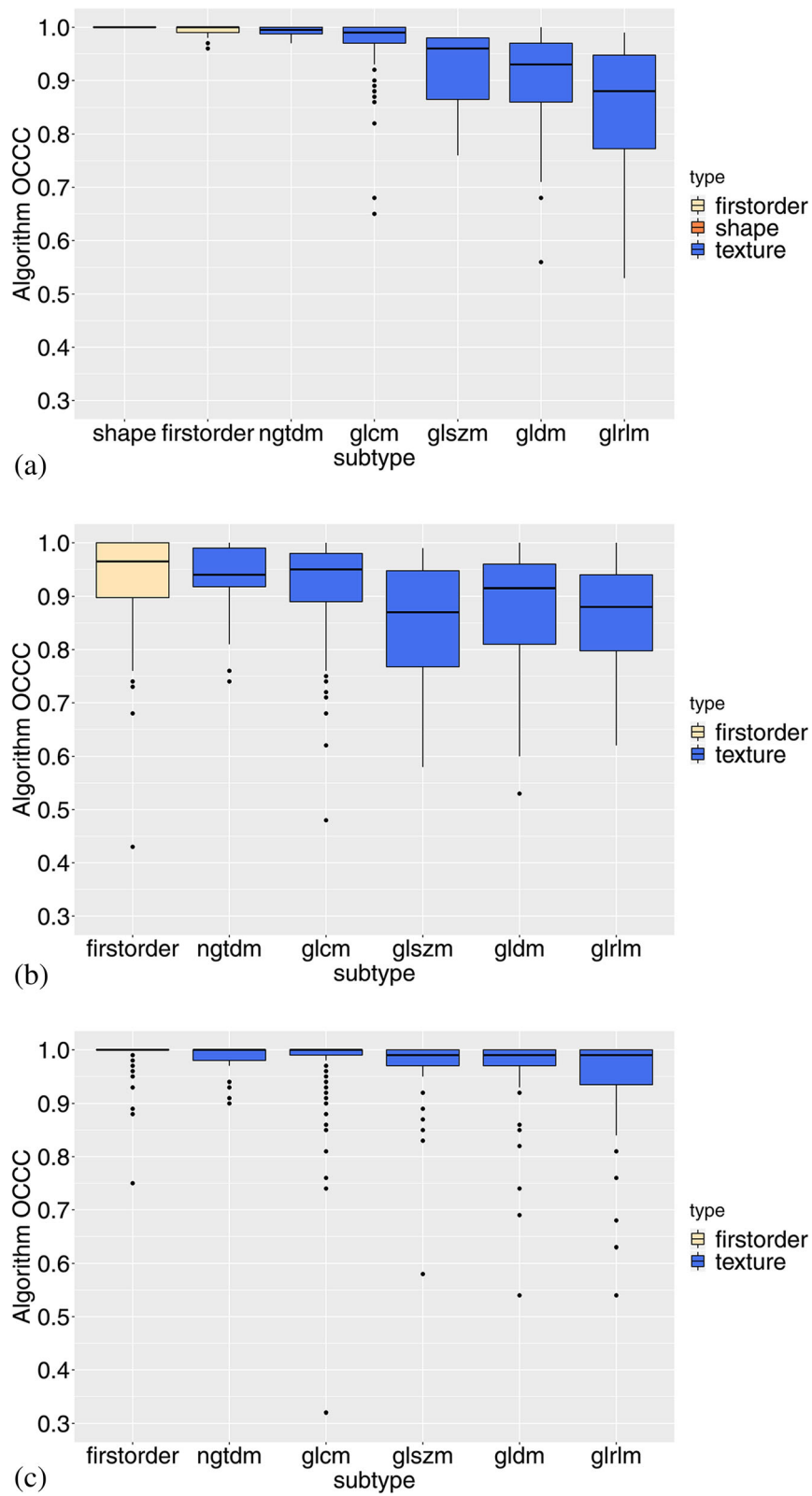


Fig. 2 Overall concordance correlation coefficient (OCCC) among the different reconstruction algorithms. The OCCC is plotted within each subtype of feature and for feature extracted from the original images (a), and the wavelet- (b) and LoG-filtered (c) images

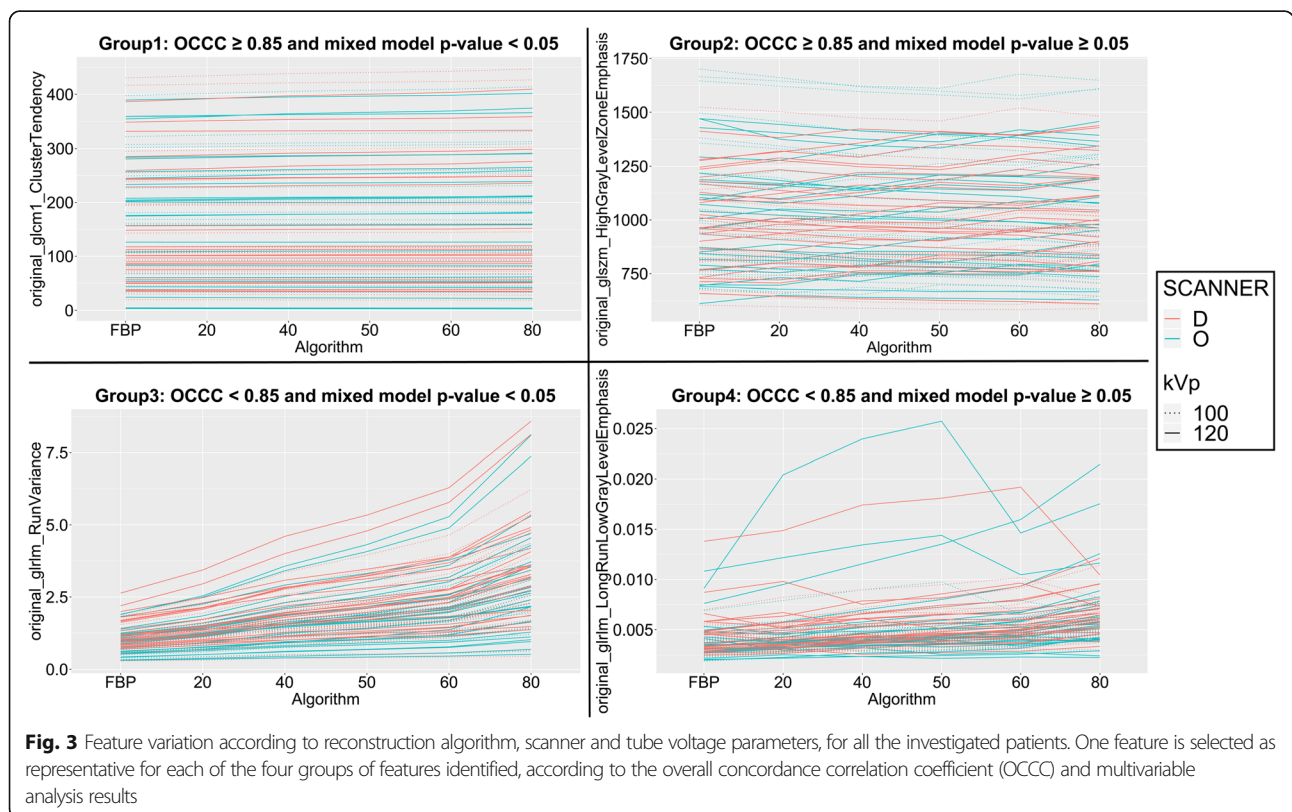
Table 3 Median OCCC values calculated for each image type and feature category

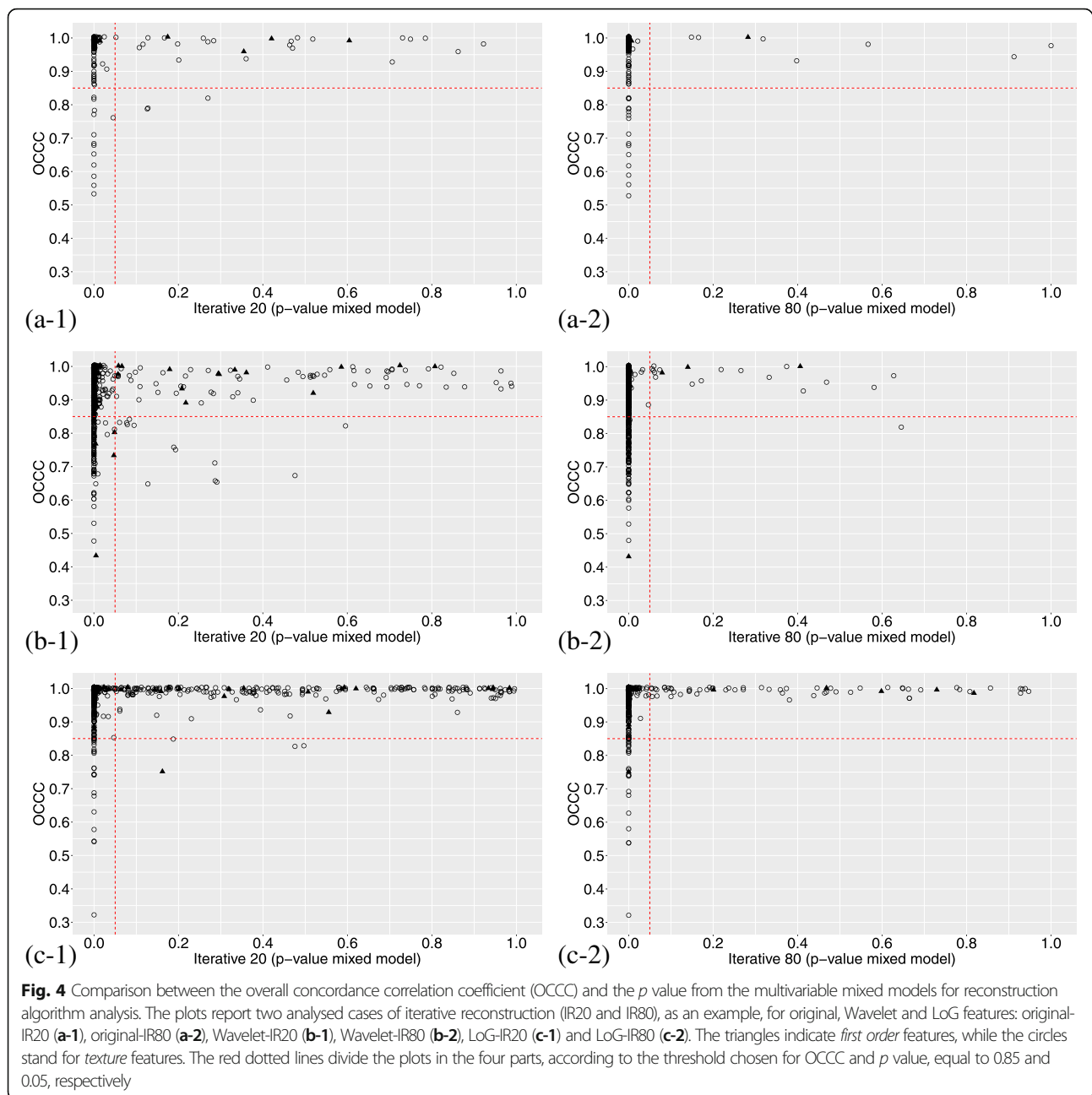
Image type	Image subtype	Feature category						
		Shape	First order	ngtdm	glcm	glszm	gldm	glrlm
Original	All	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.99 (1.00)	0.96 (0.98)	0.93 (0.97)	0.88 (0.95)
Wavelet	All	–	0.97 (0.99)	0.94 (0.97)	0.95 (0.98)	0.87 (0.94)	0.92 (0.95)	0.88 (0.95)
	LH	–	0.93 (0.98)	0.94 (0.97)	0.93 (0.98)	0.80 (0.93)	0.89 (0.94)	0.85 (0.94)
	HL	–	0.96 (0.98)	0.94 (0.98)	0.96 (0.99)	0.88 (0.94)	0.90 (0.95)	0.88 (0.95)
	HH	–	0.88 (0.96)	0.84 (0.92)	0.89 (0.95)	0.81 (0.92)	0.88 (0.95)	0.87 (0.94)
	LL	–	1.00 (1.00)	1.00 (1.00)	0.99 (1.00)	0.97 (0.99)	0.96 (0.99)	0.94 (0.97)
LoG	All	–	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.99 (1.00)	0.99 (1.00)	0.99 (1.00)
	0.5 mm	–	0.97 (0.99)	0.96 (0.98)	0.96 (0.99)	0.87 (0.95)	0.86 (0.93)	0.85 (0.93)
	1.0 mm	–	1.00 (1.00)	0.98 (0.99)	0.99 (1.00)	0.98 (0.99)	0.98 (0.99)	0.97 (0.98)
	1.5 mm	–	1.00 (1.00)	0.99 (1.00)	1.00 (1.00)	0.99 (1.00)	1.00 (1.00)	0.99 (1.00)
	2.5 mm	–	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	5.0 mm	–	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)

In parentheses, the results obtained when restricting the analysis to the settings most used in clinics (IR40, IR50, IR60 and IR80). *IR* iterative reconstruction, *glcm* grey level co-occurrence matrix, *gldm* grey level dependence matrix, *glrlm* grey level run length matrix, *glszm* grey level size zone matrix, *ngtdm* neighbouring grey tone difference matrix, *LoG* Laplacian of Gaussian, *OCCC* overall concordance correlation coefficient

statistical analyses (concordance analysis and multivariable mixed model), we showed how radiomic features can be classified in four different groups exhibiting different behaviour in relation to the reconstruction settings, which might require different selection or correction strategies to guarantee robustness and

reproducibility of radiomic results. We believe that such combined approach is useful to provide more complete information as compared to the use of one model alone, and it might allow a more comprehensive handling of the reproducibility issue.





Indeed, if considering the results of univariate concordance analysis alone, 1260/1414 features (89%) exhibiting $OCCC \geq 0.85$ would be considered as reproducible and usable, without additional correction, for a radiomic analysis performed on a clinical database with similar characteristics as the one considered here. Conversely, the remaining 154/1414 features (11%) exhibiting $OCCC < 0.85$ would be excluded. Nonetheless, the subset of such excluded features falling in group 3 ($OCCC < 0.85$ and mixed model FDR-adjusted p value < 0.05) could be retrieved and

included back in the analysis after properly accounting for the fact that their dependence on reconstruction setting is systematic among patients. The parameters needed to apply such correction are given as output by the multivariable model itself. To provide an example, we applied such correction to a feature belonging to group 3, *original_grlm_RunVariance*, and compared the values obtained for different reconstruction settings before and after the correction (Figure S4). The systematic trend observed when varying the reconstruction setting is reduced,

Table 4 Percentage of features falling in each of the four groups

Image type	Image subtype	Group 1		Group 2		Group 3		Group 4	
		IR20	IR80	IR20	IR80	IR20	IR80	IR20	IR80
Original	All (shape excluded)	70.0	82.9	18.6	5.7	9.3	11.4	2.1	0.0
Wavelet	All	65.0	75.9	14.3	3.4	18.4	20.5	2.3	0.2
	LH	16.8	19.3	2.9	0.4	4.6	5.3	0.7	0.0
	HL	18.0	20.7	4.1	1.4	2.7	2.9	0.2	0.0
	HH	10.9	13.9	3.6	0.5	9.8	10.3	0.7	0.2
	LL	19.3	22.0	3.7	1.1	1.3	2.0	0.7	0.0
LoG	All	64.6	87.6	32.3	9.3	2.7	3.1	0.4	0.0
	0.5 mm	13.4	16.0	3.6	1.0	2.6	3.0	0.4	0.0
	1.0 mm	15.5	19.0	4.4	0.9	0.1	0.1	0.0	0.0
	1.5 mm	13.1	18.4	6.9	1.6	0.0	0.0	0.0	0.0
	2.5 mm	11.6	17.6	8.4	2.4	0.0	0.0	0.0	0.0
	5.0 mm	11.0	16.6	9.0	3.4	0.0	0.0	0.0	0.0

The results are reported for the IR20 and the IR80 reconstructions. The percentage for the original images is evaluated excluding the *shape* features. *IR* iterative reconstruction, *LoG* Laplacian of Gaussian

which might allow to retain the feature (and its potential informative content) for the radiomic analysis.

Similarly, among the features exhibiting $OCCC \geq 0.85$, the subgroup falling in group 1 ($OCCC \geq 0.85$ and mixed model FDR-adjusted p value < 0.05) might require a correction before being considered reproducible. The real necessity of such correction might depend on the clinical question that the radiomic analysis is supposed to answer. For example, if the aim is to discriminate two patient populations for which the difference—in terms of radiomic features—exists but is very small, even the slight feature variation introduced by different IR blending levels may have a relevant impact, confounding the data and impairing the ability of radiomics to reach its goal. In this case, the feature correction should be applied, similarly for features in group 3, despite the $OCCC \geq 0.85$ would suggest feature reproducibility. If instead the difference between the features of two populations is far larger than the fluctuations due to the different reconstruction settings, it might be irrelevant to perform the correction or not. We plan to investigate these aspects in future studies for different clinical endpoints on the NSCLC population.

It must be noted that, in our sample the features falling in the above cited groups 1 and 3 are the vast majority, with highest prevalence in group 1 as compared to group 3 (Fig. 4 and Table 4). These features are the ones for which a trend among reconstruction blending levels has been observed and an appropriate correction may be thus applied to take into account these differences. This is of importance, because it suggests that possible differences in radiomic features according to different blending levels may be properly corrected for the majority of

the features, thus avoiding discharging them from further statistical analyses. A similar behaviour was identified by Prezzi et al., analysing CT images of 28 patients with primary colorectal cancer with multilevel linear regression [25]. They studied the impact of the reconstruction strength by applying an ASIR algorithm in steps of 20% from 0 to 100%, in a controlled acquisition setting. Similar to our results, they found that the majority of the features extracted from original images had a systematic trend (increasing or decreasing linear behaviour) with the reconstruction strength. The features belonging to group 2 ($OCCC \geq 0.85$ and p value ≥ 0.05) for all the IR blending levels can be definitively considered reproducible without need of any correction for any clinical endpoint, but in our sample their number is very small: 7 features extracted from the original images, 11 from the wavelet-filtered images and 55 from the LoG-filtered images.

Lastly, the features in group 4 ($OCCC < 0.85$ and p value ≥ 0.05) should be rejected without possibility of correction. In our sample, however, this group was poorly populated.

It should be highlighted that the data discussed so far refer to a heterogeneous database including all the six reconstruction settings from FBP to IR80. The full data reported in Tables S3, S4, S5 allow to derive conclusions for database including only FBP and a subset of the IR blending levels here considered, and the results of our subanalysis (including only IR40, IR50, IR60 and IR80) can be taken as reference for the current clinical scenario where FBP is progressively replaced by iterative algorithms. As expectable, the results of such subanalysis are quite similar to the ones of the main analysis, but with a general increase in feature reproducibility.

In addition, the result of the comparison of pairwise reconstruction algorithms for the original images (Figure S3) shows that if we changed the reference IR blending level in the multivariable model, the number of features whose value changes significantly between two IR blending levels would not vary considerably; hence, the above considerations hold valid independently on this choice. It is also important to note that, even after applying the feature selection and correction described so far, the number of reproducible features is likely to be very high, but many of them are highly correlated, so their number would be further reduced by clustering procedures (Table S1).

As possible limitations of the present study, we acknowledge the relatively small number of patients, the impossibility to investigate radiomic feature repeatability, the lack of an external validation of our results, and the inability to specifically account for the possible effect of segmentation performed by different operators. In this study, the segmentation by multiple operators should have had a negligible impact when focusing on the effect of reconstruction algorithm intra-patient, since in this case the region of interest was fixed across the different reconstruction settings. However, it might have slightly affected the assessment of inter-patients' behaviour and the analysis on scanner and tube voltage dependence. An interesting future alternative may be applying automatic approach based on deep learning. Regarding repeatability, we plan to account for this effect in future studies either on clinical images as previously performed by Zwanenburg et al. [53], or relying on dedicated phantoms under development in our group. Another limitation of our study is the use of two different iodinated-contrast media (Ultravist® 370 and Visipaque® 320). While all the patients scanned on the Discovery CT750 HD scanner received the Ultravist® 370, in the two populations scanned on the Optima 660 scanner this type of contrast was injected only in about half of the patients (the 52% and the 56% at 100 kVp and 120 kVp, respectively). The administration of two different contrast media may have affected the texture of the CT image. However, a previous study of our group performed on CT images of NSCLC patients [14] showed that the radiomic features were not significantly influenced by the different contrast media, and therefore this factor was not investigated in this study.

In conclusion, the present study confirmed that the use of different blending levels during CT reconstruction may introduce confounding factors in a radiomic analysis of NSCLC population, especially when a wide range of different blending levels are present in the dataset. Aiming to improve the robustness and efficacy of radiomic studies, a novel approach for the identification of reproducible features in a given dataset is proposed, to be

applied before redundancy reduction and correlation analysis with clinical endpoints.

Abbreviations

ASIR: Adaptive statistical iterative reconstruction; CT: Computed tomography; FBP: Filtered backprojection; FDR: False discovery rate; GLCM: Grey level co-occurrence matrix; GLDM: Grey level dependence matrix; GLRLM: Grey level run length matrix; GLSZM: Grey level size zone matrix; HU: Hounsfield unit; IR: Iterative reconstruction; LoG: Laplacian of Gaussian; NGTD M: Neighbouring grey tone difference matrix; NSCLC: Non-small-cell lung cancer; OCCC: Overall concordance correlation coefficient

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41747-021-00258-6>.

Additional file 1: Supplementary Methods. Details on the extracted radiomic features. **Table S1.** List of radiomic features included in the study; the name of the feature and the corresponding category, sub-category and cluster is reported. **Table S2.** False Discovery Rate (FDR) adjusted p values for univariate and multivariable analysis for the effect of scanner and tube voltage on selected radiomic features. **Table S3.** OCCC values and False Discovery Rate (FDR) adjusted p values from the multivariable mixed model for the reconstruction algorithm impact for all the original features (shape features are not included because their value was always 1 since the same VOI was used for all the reconstructions). **Table S4.** OCCC values and False Discovery Rate (FDR) adjusted p value from the multivariable mixed model for the reconstruction algorithm impact for all the wavelet features. **Table S5.** OCCC values and False Discovery Rate (FDR) adjusted p value from the multivariable mixed model for the reconstruction algorithm impact (FDR corrected) for all the LoG features. **Table S6.** Percentage of features falling in each of the 4 groups, for the IR50 and the IR80 reconstructions. The results reported in this table refer to the sub-analysis performed on IR40, IR50, IR60 and IR80 (taking IR40 for comparison), similarly to Table 4 for the complete analysis. The percentage for the original images is evaluated excluding the features of the shape category. **Figure S1.** Study flowchart with exclusion criteria. **Figure S2.** Overall Concordance Correlation Coefficient (OCCC) for concordance between different algorithms for the sub-analysis of the blending levels mostly used in our clinical setting (IR40, IR50, IR60 and IR80). The OCCC is plotted within each subtype of feature and for feature extracted from the original images (a), and the Wavelet (b) and LoG-filtered (c) images. **Figure S3.** Heatmap representing the number of features significantly different for paired comparisons of reconstruction algorithm strength (Wilcoxon signed rank-test), for the original images. Shape feature excluded. **Figure S4.** Box plot representing the distribution of the radiomic features (a) *original_glrlm_RunVariance* and (b) modified *original_glrlm_RunVariance*, according to algorithm. The modified *original_glrlm_RunVariance* was obtained after rescaling, using the correspondent coefficients for each algorithm obtained in the mixed model. Minimum and maximum are depicted by whiskers, the box signifies the upper and lower quartiles, the median and the mean are represented, respectively by a line and a small rhombus within the box.

Acknowledgements

LR, AL, and MM acknowledge the AIM-INFN project. The authors acknowledge the support of the research grants from the Italian Ministry of Health (GR-2016-02362050 and RCR-ACC-2019-WG12).

Authors' contributions

DO, and FB equally contributed to the conception and the design of the work; SRa, SDA, LR, FB, and CF to the preparation and analysis of data. SRa, LR, SDA, FB and DO to the interpretation of data. LR, FB, and SRa have drafted the work and substantively revised it. SRi, SDA, AL, MM, CF, CR, MC, DO, and RO supported in the interpretation and reviewing. All authors have read and approved the final version of the manuscript.

Funding

The work was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds. LR is supported by a research grant from the Italian Ministry of Health, GR-2016-02362050. The APC was funded by a research grant from the Italian Ministry of Health, RCR-ACC-2019-WG12.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of European Institute of Oncology (protocol code UID 2412, 11/05/2020).

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Radiation Research Unit, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy. ²Department of Physics, Università degli Studi di Pavia and INFN, via Bassi 6, 27100 Pavia, Italy. ³Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy. ⁴Clinica di Radiologia EOC, Istituto Imaging della Svizzera Italiana (IIMSI), via Tesserete 46, 6900 Lugano, Switzerland. ⁵Università della Svizzera italiana, via G.Buffi 13, 6900 Lugano, Switzerland. ⁶Postgraduate School of Diagnostic and Interventional Radiology, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milan, Italy. ⁷Department of Radiology, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy. ⁸Scientific Directorate, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy. ⁹Medical Physics Unit, IEO European Institute of Oncology IRCCS, via Ripamonti 435, 20141 Milan, Italy.

Received: 20 September 2021 Accepted: 16 December 2021

Published online: 25 January 2022

References

- Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebbers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 5:4006. <https://doi.org/10.1038/ncomms5006>
- Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, van Wijk Y, Woodruff H, van Soest J, Lustberg T, Roelofs E, van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 14:749–762. <https://doi.org/10.1038/nrclinonc.2017.141>
- Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ (2012) QIN “radiomics: the process and the challenges.”. *Magn Reson Imaging* 30:1234–1248. <https://doi.org/10.1016/j.mri.2012.06.010>
- Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 278:563–577. <https://doi.org/10.1148/radiol.2015151169>
- Avanzo M, Stancanello J, El Naqa I (2017) Beyond imaging: the promise of radiomics. *Phys Medica PM Int J Devoted Appl Phys Med Biol Off J Ital Assoc Biomed Phys AIFB* 38:122–139. <https://doi.org/10.1016/j.ejmp.2017.05.071>
- Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys* 102:1143–1158. <https://doi.org/10.1016/j.ijrobp.2018.05.053>
- Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, Huang SH, Purdie TG, O’Sullivan B, Aerts HJWL, Jaffray DA (2019) Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol J Eur Soc Ther Radiol Oncol* 130:2–9. <https://doi.org/10.1016/j.radonc.2018.10.027>
- Fornacon-Wood I, Faivre-Finn C, O’Connor JPB, Price GJ (2020) Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. *Lung Cancer* 146:197–208. <https://doi.org/10.1016/j.lungcan.2020.05.028>
- Espinasse M, Pitre-Champagnat S, Charmettant B, Bidault F, Volk A, Balleyguier C, Lassau N, Caramella C (2020) CT texture analysis challenges: influence of acquisition and reconstruction parameters: a comprehensive review. *Diagn Basel Switz* 10. <https://doi.org/10.3390/diagnostics10050258>
- Robins M, Solomon J, Hoye J, Abadi E, Marin D, Samei E (2019) Systematic analysis of bias and variability of texture measurements in computed tomography. *J Med Imaging* 6. <https://doi.org/10.1117/1.JMI.6.3.033503>
- van Timmeren JE, Cester D, Tanadini-Lang S, Alkadi H, Baessler B (2020) Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights Imaging* 11:91. <https://doi.org/10.1186/s13244-020-00887-2>
- Nair JKR, Saeed UA, McDougall CC, Sabri A, Kovacina B, Raidu BVS, Khokhar RA, Probst S, Hirsh V, Chankowsky J, van Kempen L, Taylor J (2020) Radiogenomic models using machine learning techniques to predict EGFR mutations in non-small cell lung cancer. *Can Assoc Radiol J J Assoc Can Radiol* 846537119899526:109–119. <https://doi.org/10.1177/0846537119899526>
- Shiri I, Maleki H, Hajianfar G, Abdollahi H, Ashrafinia S, Hatt M, Zaidi H, Oveisi M, Rahmim A (2020) Next-generation radiogenomics sequencing for prediction of EGFR and KRAS mutation status in NSCLC patients using multimodal imaging and machine learning algorithms. *Mol Imaging Biol* 22: 1132–1148. <https://doi.org/10.1007/s11307-020-01487-8>
- Botta F, Raimondi S, Rinaldi L, Bellerba F, Corso F, Bagnardi V, Origgi D, Minelli R, Pitoni G, Petrella F, Spaggiari L, Morganti AG, del Grande F, Bellomi M, Rizzo S (2020) Association of a CT-based clinical and radiomics score of non-small cell lung cancer (NSCLC) with lymph node status and overall survival. *Cancers* 12. <https://doi.org/10.3390/cancers12061432>
- Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I (2018) A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med Off Publ Soc Nucl Med* 59:1321–1328. <https://doi.org/10.2967/jnumed.117.199935>
- Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* 291:53–59. <https://doi.org/10.1148/radiol.2019182023>
- 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* 65:015010. <https://doi.org/10.1088/1361-6560/ab6177>
- Fave X, Cook M, Frederick A, Zhang L, Yang J, Fried D, Stingo F, Court L (2015) Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imaging Graph Off J Comput Med Imaging Soc* 44:54–61. <https://doi.org/10.1016/j.compmedimag.2015.04.006>
- Lu L, Ehmke RC, Schwartz LH, Zhao B (2016) Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS One* 11: e0166550. <https://doi.org/10.1371/journal.pone.0166550>
- Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, Schwartz LH (2016) Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep* 6: 23428. <https://doi.org/10.1038/srep23428>
- Park S, Lee SM, Do K-H, Lee JG, Bae W, Park H, Jung KH, Seo JB (2019) Deep learning algorithm for reducing CT slice thickness: effect on reproducibility of radiomic features in lung cancer. *Korean J Radiol* 20:1431–1440. <https://doi.org/10.3348/kjr.2019.0212>
- Erdal BS, Demirel M, Little KJ, Amadi CC, Ibrahim GFM, O’Donnell TP, Grimmer R, Gupta V, Prevedello LM, White RD (2020) Are quantitative features of lung nodules reproducible at different CT acquisition and reconstruction parameters? *PLoS One* 15:e0240184. <https://doi.org/10.1371/journal.pone.0240184>
- Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, Hwang EJ, Goo JM (2016) Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One* 11:e0164924. <https://doi.org/10.1371/journal.pone.0164924>
- Solomon J, Mileto A, Nelson RC, Roy Choudhury K, Samei E (2016) Quantitative features of liver lesions, lung nodules, and renal stones at

- multi-detector row CT examinations: dependency on radiation dose and reconstruction algorithm. *Radiology* 279:185–194. <https://doi.org/10.1148/radiol.2015150892>
25. Prezzi D, Owczarczyk K, Bassett P, Siddique M, Breen DJ, Cook GJR, Goh V (2019) Adaptive statistical iterative reconstruction (ASIR) affects CT radiomics quantification in primary colorectal cancer. *Eur Radiol* 29:5227–5235. <https://doi.org/10.1007/s00330-019-06073-3>
26. Sung P, Lee JM, Joo I, Lee S, Kim TH, Ganeshan B (2019) Evaluation of the impact of iterative reconstruction algorithms on computed tomography texture features of the liver parenchyma using the filtration-histogram method. *Korean J Radiol* 20:558–568. <https://doi.org/10.3348/kjr.2018.0368>
27. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, Mitra S, Shankar BU, Kikinis R, Haibe-Kains B, Lambin P, Aerts HJWL (2014) Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* 9:e102107. <https://doi.org/10.1371/journal.pone.0102107>
28. Owens CA, Peterson CB, Tang C, Koay EJ, Yu W, Mackin DS, Li J, Salehpour MR, Fuentes DT, Court LE, Yang J (2018) Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS One* 13:e0205003. <https://doi.org/10.1371/journal.pone.0205003>
29. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, Huellner M, Opitz I, Weder W, Frauenfelder T, Guckenberger M, Tanadini-Lang S (2018) Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol Stockh Swed* 57:1070–1074. <https://doi.org/10.1080/0284186X.2018.1445283>
30. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Court LE (2016) Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. *Transl Cancer Res* 5:349–363. <https://doi.org/10.21037/78709>
31. Shafiq-UI-Hassan M, Latifi K, Zhang G et al (2018) Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep* 8:10545. <https://doi.org/10.1038/s41598-018-28895-9>
32. Wang HCY, Donovan EM, Nisbet A, South CP, Alobaidli S, Ezhil V, Phillips I, Prakash V, Ferreira M, Webster P, Evans PM (2019) The stability of imaging biomarkers in radiomics: a framework for evaluation. *Phys Med Biol* 64:165012. <https://doi.org/10.1088/1361-6560/ab23a7>
33. Park BW, Kim JK, Heo C, Park KJ (2020) Reliability of CT radiomic features reflecting tumour heterogeneity according to image quality and image processing parameters. *Sci Rep* 10:3852. <https://doi.org/10.1038/s41598-020-60868-9>
34. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77:e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
35. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrin L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desserot MC, Dinapoli N, Dinh CV, Echeagaray S, el Naqa I, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkiewicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsma NM, Socarras Fernandez J, Spezi E, Steenbakkens RJHM, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhyaya T, Valentini V, van Dijk LV, van Griethuysen J, van Velden FHP, Whybra P, Richter C, Lööck S (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295:328–338. <https://doi.org/10.1148/radiol.2020.91145>
36. Shafiq-UI-Hassan M, Zhang GG, Latifi K et al (2017) Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys* 44:1050–1062. <https://doi.org/10.1002/mp.12123>
37. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, Court L (2017) Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One* 12:e0178524. <https://doi.org/10.1371/journal.pone.0178524>
38. Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, Dekker A, Fijten R, Monshouwer R (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>
39. Kakino R, Nakamura M, Mitsuyoshi T, Shintani T, Hirashima H, Matsuo Y, Mizowaki T (2020) Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. *Phys Medica PM Int J Devoted Appl Phys Med Biol Off J Ital Assoc Biomed Phys AIFB* 69:176–182. <https://doi.org/10.1016/j.ejmp.2019.12.019>
40. Lee G, Gommers R, Waselewski F, Wohlfahrt K, O’Leary A (2019) PyWavelets: a Python package for wavelet analysis. *J Open Source Softw* 4:1237. <https://doi.org/10.21105/joss.01237>
41. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, Lambin P, Haibe-Kains B, Mak RH, Aerts HJWL (2015) CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol J Eur Soc Ther Radiol Oncol* 114:345–350. <https://doi.org/10.1016/j.radonc.2015.02.015>
42. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z (2016) Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Sci Rep* 6:34921. <https://doi.org/10.1038/srep34921>
43. Ferreira Junior JR, Koenigkam-Santos M, Cipriano FEG, Fabro AT, Azevedo-Marques PM (2018) Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Comput Methods Programs Biomed* 159:23–30. <https://doi.org/10.1016/j.cmpb.2018.02.015>
44. Barnhart HX, Haber M, Song J (2002) Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 58:1020–1027. <https://doi.org/10.1111/j.0006-341x.2002.01020.x>
45. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, Lambin P (2016) Test-retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomogr Ann Arbor Mich* 2:361–365. <https://doi.org/10.18383/j.tom.2016.00208>
46. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, Sosef MN, Raat FHPJ, van der Zande FHR, Das M, van Elmpt W, Lambin P (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 Stockh Swed* 56:1544–1553. <https://doi.org/10.1080/0284186X.2017.1351624>
47. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
48. R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria Retrieved from <https://www.R-project.org/>
49. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones AK, Court L (2015) Measuring computed tomography scanner variability of radiomics features. *Invest Radiol* 50:757–765. <https://doi.org/10.1097/RLI.0000000000000180>
50. Buch K, Li B, Qureshi MM, Kuno H, Anderson SW, Sakai O (2017) Quantitative assessment of variation in CT parameters on texture features: pilot study using a nonanatomic phantom. *AJNR Am J Neuroradiol* 38:981–985. <https://doi.org/10.3174/ajnr.A5139>
51. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J et al (2018) Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology* 288:407–415. <https://doi.org/10.1148/radiol.2018172361>
52. Zheng Y, Solomon J, Choudhury K, Marin D, Samei E (2017) Accuracy and variability of texture-based radiomics features of lung lesions across CT imaging conditions. *Proc SPIE, Medical Imaging 2017: Physics of Medical Imaging*, 10132:1397–403. <https://doi.org/10.1117/12.2255806>
53. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, Lööck S (2019) Assessing robustness of radiomic features by image perturbation. *Sci Rep* 9:614. <https://doi.org/10.1038/s41598-018-36938-4>

Publisher’s Note

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