Diagn Interv Radiol 2024; DOI: 10.4274/dir.2023.232543

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Influence of image preprocessing on the segmentation-based reproducibility of radiomic features: *in vivo* experiments on discretization and resampling parameters

Burak Koçak
Sabahattin Yüzkan
Samet Mutlu
Mehmet Karagülle
Ahmet Kala
Mehmet Kadıoğlu
Sıla Solak
Şeyma Sunman
Zişan Hayriye Temiz
Ali Kürşad Ganiyusufoğlu

University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

To systematically investigate the impact of image preprocessing parameters on the segmentation-based reproducibility of magnetic resonance imaging (MRI) radiomic features.

METHODS

The MRI scans of 50 patients were included from the multi-institutional Brain Tumor Segmentation 2021 public glioma dataset. Whole tumor volumes were manually segmented by two independent readers, with the participation of eight readers. Radiomic features were extracted from two sequences: T2-weighted (T2) and contrast-enhanced T1-weighted (T1ce). Two methods were considered for discretization: bin count (i.e., relative discretization) and bin width (i.e., absolute discretization). Ten discretization (five for each method) and five resampling parameters were varied while other parameters were fixed. The intraclass correlation coefficient (ICC) was used for reliability analysis based on two commonly used cut-off values (0.75 and 0.90).

RESULTS

Image preprocessing parameters had a significant impact on the segmentation-based reproducibility of radiomic features. The bin width method yielded more reproducible features than the bin count method. In discretization experiments using the bin width on both sequences, according to the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features ranged from 70% to 84% and from 35% to 57%, respectively, with an increasing percentage trend as parameter values decreased (from 84 to 5 for T2; 100 to 6 for T1ce). In the resampling experiments, these ranged from 53% to 74% and from 10% to 20%, respectively, with an increasing percentage trend from lower to higher parameter values (physical voxel size; from 1 x 1 x 1 to 2 x 2 x 2 mm^3).

CONCLUSION

The segmentation-based reproducibility of radiomic features appears to be substantially influenced by discretization and resampling parameters. Our findings indicate that the bin width method should be used for discretization and lower bin width and higher resampling values should be used to allow more reproducible features.

KEYWORDS

Reproducibility, preprocessing, radiomics, texture analysis, biomarker

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 04 October 2023; revision requested 01 November 2023; accepted 14 November 2023.

Epub: 11.12.2023

Publication date: 13.05.2024

DOI: 10.4274/dir.2023.232543

Radiomics is a field of medical image analysis that enables the digital decoding of images into high-throughput quantitative features.[1] Medical images may contain hidden patterns, indicating the underlying pathophysiology of the examined tissue. Based on this assumption, radiomic features derived from these images might help characterize tissues and guide clinical decision-making.[1,2] Support for this notion has arisen from numerous studies that have addressed the capability of radiomics in making predictions regarding different clinical endpoints.[3] There has been an exponential increase in publications related to

You may cite this article as: Koçak B, Yüzkan S, Mutlu S, et al. Influence of image preprocessing on the segmentation-based reproducibility of radiomic features: *in vivo* experiments on discretization and resampling parameters. *Diagn Interv Radiol*. 2024;30(3):152-162.

radiomics, with a yearly growth rate of 19.6% and a doubling time of 3.9 years.[4] However, reproducing and validating published studies is still challenging due to a lack of standardized definitions, parameter settings, and inadequate reporting.[5-9]

Before implementing radiomics in clinical practice, it is necessary to have a thorough understanding of the reproducibility of radiomic features. Many previous publications have emphasized the dependency of radiomic features on different factors, such as temporal variability,[10,11] scanning parameters,[12-14] delineation uncertainty,[15,16] reconstruction algorithms,[17] preprocessing,[8] and organ motion.[18] The absolute value and statistical distribution of the radiomics features are considerably affected by the aforementioned determinants, which in turn affects the robustness and generalizability of any subsequent analysis derived from these features. To overcome this divergence, the Image Biomarker Standardization Initiative (IBSI) attempted to standardize the radiomic feature extraction process, focusing on the issues of repeatability, reproducibility, and validation in quantitative image analysis and radiomics.[5] According to this initiative, standardized image processing should be performed before radiomic feature extraction.[5] Nonetheless, no specific processing parameter settings have been published to date, which underlines the requirement for additional research.[8,19,20]

One of the most important steps in the radiomic pipeline that affects reproducibility is segmentation or delineation.[21,22] For example, a feature might be highly reproducible in a test–retest setting, but there is no guarantee that this feature will be robust after segmentation. Segmentation-based reproducibility analysis is extensively used to reduce the high dimensionality of radiomics data as a data handling step for subsequent predictive modeling procedures.[2,23] However, only a limited number of studies have focused on the impact of preprocessing settings on segmentation-based feature reproducibility.[24,25] Duron et al.[24] studied magnetic resonance imaging (MRI)-based radiomic features of lachrymal gland tumors and breast lesions with a focus on discretization techniques. Lu et al.[25] investigated positron emission tomography/computed tomography (PET/CT)-based radiomic features in patients with nasopharyngeal carcinoma, again with a focus on discretization. No research has specifically assessed the impact of both image voxel resampling and gray-level discretization on the segmentation-based reproducibility of the radiomic features. However, these two preprocessing methods are frequently encountered in radiomic feature extraction software tools.

The purpose of this study was to systematically investigate the effect of image preprocessing parameters on the segmentation-based reproducibility of radiomic features from MRI and to recommend reasonable parameter settings for achieving highly reproducible features.

Main points

- Variations of image preprocessing parameters, regarding discretization and resampling, have a significant impact on the segmentation-based reproducibility of radiomic features.
- The bin width method yields more reproducible features than the bin count method for discretization.
- Using lower bin width values and higher resampling values could help produce more reproducible features.
- The optimal preprocessing parameters should be determined within the radiomic pipeline.
- To allow replication, preprocessing parameters should be transparently reported in radiomic publications due to their importance.

Methods

Figure 1 depicts the key study steps to help readers understand the methodology.

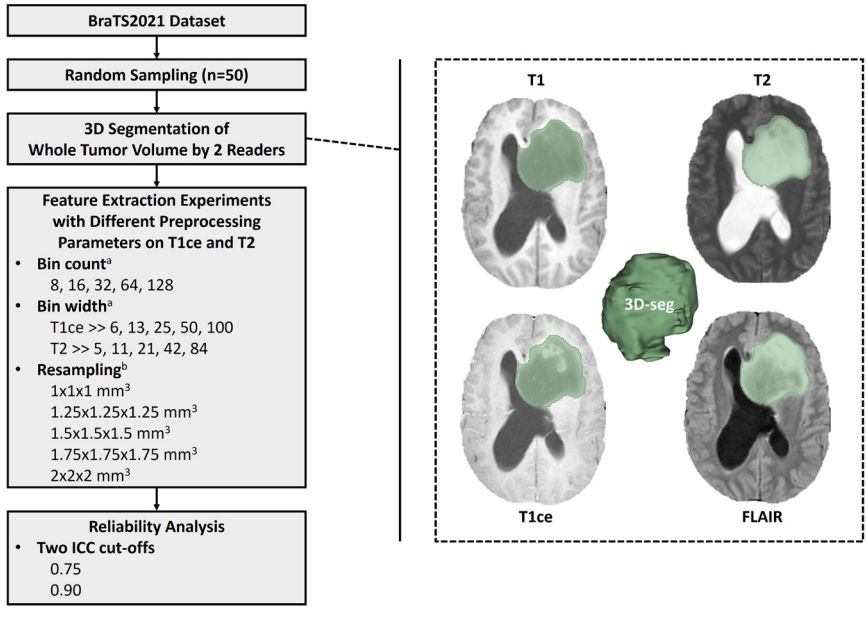


Figure 1. Key study steps and segmentation approach. 3D, three-dimensional; ICC, intraclass correlation coefficient; T1, T1-weighted; T2, T2-weighted; T1ce, contrast-enhanced T1-weighted; FLAIR, fluid-attenuated inversion recovery; 3D-seg, three-dimensional segmentation. [a]Resampling fixed to 1 x 1 x 1 mm³. [b]Discretization fixed to a bin count of 32.

Dataset

In this study, we used the Brain Tumor Segmentation (BraTS) 2021 public glioma dataset,[26-28] which does not require local ethical approval. The MRI data for the BraTS 2021 challenge were collected using various clinical protocols and scanners from a variety of data-contributing institutions. There were four MRI sequences in the dataset: T1-weighted (T1), T2-weighted (T2), contrast-enhanced T1-weighted (T1ce), and fluid-attenuated inversion recovery (FLAIR). All BraTS MRI scans underwent standardized preprocessing, which included the conversion of Digital Imaging and Communications in Medicine-format files to Neuroimaging Informatics Technology Initiative format, co-registration to the same anatomical template (SRI24),[29] isotropic voxel resampling (1 x 1 x 1 mm³), and skull-stripping.[30]

For this reproducibility study, 50 patients with gliomas were randomly selected. Patient identifiers are provided in the Supplementary Table S1. Readers who performed the segmentation used all four sequences. Only two sequences-T2 and T1ce-were used for the preprocessing experiments to assess the dependency of the results on the different sequences; the use of more sequences may have become unfeasible considering the workload and complexity of the study. The T2 sequence was selected to represent the outermost boundary of the tumor, and T1ce was used to evaluate the radiomic features on a different image contrast, considering the relatively homogeneous appearance of glial tumors in T2 compared with T1ce.

Segmentation

The glial tumors were manually segmented using 3D Slicer software v4.11. The pathological high signal intensity that appears in T2 and FLAIR sequences was used to segment the entire tumor volume. Readers were also free to use any of the four sequences available in the dataset to determine tumor borders (T1, T2, FLAIR, and T1ce). Figure 1 also illustrates the segmentation approach.

The segmentation process involved eight readers (three radiologists and five radiology residents), with two readers (one radiology specialist and one radiology resident) for each patient. All of the specialists worked in the neuroradiology division. Two of these had ≥3 years and one had ≥1 years of experience in neuroimaging as a specialist. During the study, all of the residents were in their second or third year in radiology and on their first neuroradiology rotation.

Preprocessing

All images were normalized to a scale of 100 based on the mean and standard deviation (SD) of voxel intensity values. To avoid negative values, the voxel arrays were shifted by 300.

Experiments were conducted by changing the discretization and resampling parameters. For discretization, two methods were considered: bin count (i.e., relative discretization) and bin width (i.e., absolute discretization). The following preprocessing parameters were used for bin count: 8, 16, 32, 64, and 128. For the bin width method, the following preprocessing settings were used for T1ce: 6, 13, 25, 50, and 100; for T2: 5, 11, 21, 42, and 84. The bin width values were determined based on the first-order range in the dataset to get an approximately equal number of gray levels compared with the bin count approach. When experimenting with the above-mentioned two discretization approaches, the resampling parameter was fixed to 1 x 1 x 1 mm^3. For resampling, the physical voxel sizes were rescaled to 1 x 1 x 1, 1.25 x 1.25 x 1.25, 1.5 x 1.5 x 1.5, 1.75 x 1.75 x 1.75, and 2 x 2 x 2 mm^3. When performing the resampling experiments, the discretization parameter was fixed to a bin count of 32.

Feature extraction

Three-dimensional radiomic features, including shape and texture, were extracted in batch mode using the PyRadiomics open-source software environment (PyRadiomics v3.0.1; NumPy v1.23.5; SimpleITK v2.3.0; PyWavelet v1.4.1; Python 3.10.12).[31] The total number of features in each sequence was 1.106. Original, Laplacian of Gaussian (LoG)-filtered, and wavelet-transformed images were used in the feature extraction. The LoG filtering was performed with sigma values of 2, 4, and 6 mm, corresponding to fine, medium, and coarse patterns. The main feature classes were shape, first order, gray-level co-occurrence matrix, gray-level size zone matrix, gray-level run-length matrix, gray-level dependence matrix, and neighboring gray-tone difference matrix.

Statistical analysis

The R v4.3 (rstatix v0.7.2) and Python v3.7 (pingouin v0.5.2) software packages were utilized to conduct statistical analyses. To measure feature reproducibility, the intraclass correlation coefficient (ICC) was estimated based on two-way random effects, absolute agreement, and single measurement, under the Shrout and Fleiss convention.[32] The interpretation scale for the ICC was as follows: ICC <0.50, poor; 0.50≤ ICC <0.75, moderate; 0.75≤ ICC <0.90, good; and ICC ≥0.90, excellent.[33] Two thresholds-0.75 and 0.90-were used to report the percentages of reproducible features. The normality of the ICC values was determined using the Shapiro–Wilk test. Depending on the group distributions, paired tests, notably the one-way repeated measures analysis of variance (ANOVA) and the student t-test, were used to evaluate statistical differences in continuous variables for all and pair-wise comparisons, respectively. McNemar's test was utilized to compare the distribution of categorical variables (i.e., reproducible vs. non-reproducible features based on ICC cut-off values). Statistical results were considered significant if *P* values were ≤0.05. Multiple comparisons were subjected to multiplicity correction using the Tukey test or Bonferroni correction as appropriate. In these comparisons, statistical significance was determined based on adjusted or unadjusted but corrected *P* values, for the Tukey test and Bonferroni correction, respectively.

Results

Figure 2 presents the distribution of the ICC estimates for various preprocessing processes, including discretization with bin count, discretization with bin width, and voxel resampling. Detailed descriptive statistics of the ICC estimates based on preprocessing processes are presented in Table 1. As the bin width was reduced in the experiments, the mean ICC values increased. In experiments involving the bin count, the mean ICC val-

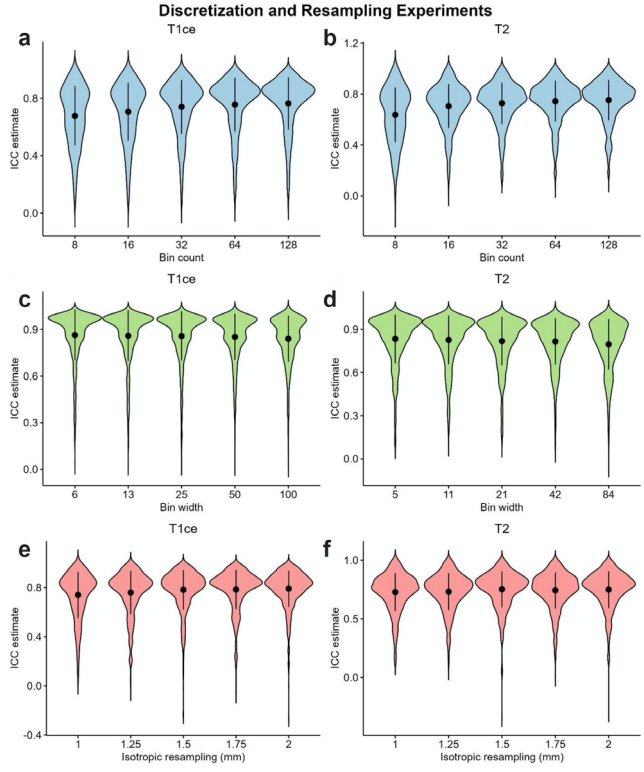


Figure 2. Distribution of the intraclass correlation coefficient (ICC) estimates for different preprocessing steps. Experiments with bin count **(a, b)**, bin width **(c, d)**, and voxel resampling **(e, f)** on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. The filled circle and bar inside the violin represent the mean and standard deviation, respectively.

ues increased as the bin count increased. Both tests revealed that an increase in the number of gray levels led to an increase in the mean ICC values and, in turn, the segmentation-based reproducibility of radiomic features. The mean ICC values were statistically significantly different and higher in the bin width group (for T1ce, mean ± SD, 0.855 ± 0.158; for T2, mean ± SD, 0.818 ± 0.169) than in the bin count group (for T1ce, mean ± SD, 0.729 ± 0.196; for T2, mean ± SD, 0.713 ± 0.180) on both of the T1ce [$t(2,764) = -28.2$, $P < 0.001$] and T2 [$t(2,764) = -22.3$, $P < 0.001$] sequences. For the resampling, the mean ICC values improved as the resampled physical voxel size increased.

Table 1. Descriptive statistics of intraclass correlation coefficients for preprocessing experiments

Sequence	Preprocessing method	Preprocessing parameter	ICC estimate	
			Mean	SD
T1ce	Bin count	8	0.678	0.206
		16	0.707	0.203
		32	0.740	0.188
		64	0.757	0.187
		128	0.765	0.182
	Bin width	6	0.864*	0.165
		13	0.859*	0.165
		25	0.858*	0.160
		50	0.852*	0.149
		100	0.840*	0.149
	Resampling	1 x 1 x 1 mm^3	0.740	0.188
		1.25 x 1.25 x 1.25 mm^3	0.760	0.178
		1.5 x 1.5 x 1.5 mm^3	0.782	0.161
		1.75 x 1.75 x 1.75 mm^3	0.784	0.160
		2 x 2 x 2 mm^3	0.791	0.147
T2	Bin count	8	0.637	0.216
		16	0.705	0.172
		32	0.728	0.163
		64	0.743	0.160
		128	0.752	0.158
	Bin width	5	0.834*	0.170
		11	0.826*	0.169
		21	0.818*	0.170
		42	0.816*	0.160
		84	0.796*	0.175
	Resampling	1 x 1 x 1 mm^3	0.728	0.163
		1.25 x 1.25 x 1.25 mm^3	0.731	0.157
		1.5 x 1.5 x 1.5 mm^3	0.752	0.154
		1.75 x 1.75 x 1.75 mm^3	0.743	0.155
		2 x 2 x 2 mm^3	0.749	0.157

*Top 10 values. ICC, intraclass correlation coefficient; SD, standard deviation; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted.

Table 2 presents the ANOVA findings for parameter differences across experimental groups. Although the effect sizes were minor (range: 0.002–0.029), all comparisons for all three preprocessing experiments were statistically significant ($P < 0.001$ for all experiments in both sequences). The statistically significant pairs following the post-hoc Tukey test are summarized in Table 3. Considering all evaluations based on sequence and preprocessing experiments, there were statistically significant differences at least between all minimum and maximum numeric values of the preprocessing parameters (e.g., bin count of 8 vs. 128; resampling 1 x 1 x 1 vs. 2 x 2 x 2 mm^3).

Figures 3 and 4 depict the percentages of features with good and excellent reproducibility in the discretization and resampling experiments, based on two typical ICC cut-off values (0.75 and 0.90). In the discretization experiments with bin count on both sequences, taking the ICC cut-offs of 0.75 and 0.90 into account, the rate of reproducible features was 36%–69% and 9%–19%, respectively, with an increasing percentage trend from lower parameter values to higher parameter values. In the discretization experiments with bin width on two sequences, with the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features was 70%–84% and 35%–57%, respectively, with an increasing percentage trend as parameter values decreased. In resampling experiments on both sequences, with the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features was 53%–74% and 10%–20%, respectively, with an increasing percentage trend from lower to higher parameter values.

Given a fixed first-order range in a sequence calculated based on the dataset, the bin width experiments outperformed the re-

Table 2. One-way repeated measures analysis of variance results of intraclass correlation coefficients

Preprocessing method	Sequence	dfN	dfD	F	Generalized eta-squared	*P*
Bin width	T1ce	1	5528	14.722	0.003	<0.001
	T2	1	5528	29.810	0.005	<0.001
Bin count	T1ce	1	5528	105.387	0.019	<0.001
	T2	1	5528	163.082	0.029	<0.001
Resampling	T1ce	1	5528	62.761	0.011	<0.001
	T2	1	5528	12.780	0.002	<0.001

T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; dfN, degrees of freedom in the numerator; dfD, degrees of freedom in the denominator.

Table 3. Statistically significant pairs after post-hoc Tukey test for one-way repeated measures analysis of variance

Preprocessing method	Sequence	Preprocessing parameters		Estimate	95% CI lower	95% CI upper	Adjusted *P*
		Group#1	Group#2				
Bin width	T1ce	6	100	−0.024	−0.042	−0.006	0.004
		13	100	−0.019	−0.038	−0.001	0.032
	T2	5	84	−0.038	−0.058	−0.018	<0.001
		11	84	−0.030	−0.050	−0.011	<0.001
		21	84	−0.022	−0.042	−0.003	0.016
		42	84	−0.020	−0.040	<0.001	0.044
Bin count	T1ce	8	16	0.029	0.006	0.051	0.005
		8	32	0.062	0.040	0.085	<0.001
		8	64	0.079	0.056	0.101	<0.001
		8	128	0.087	0.064	0.109	<0.001
		16	32	0.034	0.011	0.056	<0.001
		16	64	0.050	0.028	0.072	<0.001
		16	128	0.058	0.035	0.080	<0.001
		32	128	0.024	0.002	0.047	0.028
	T2	8	16	0.068	0.047	0.088	<0.001
		8	32	0.091	0.071	0.111	<0.001
		8	64	0.106	0.086	0.126	<0.001
		8	128	0.115	0.094	0.135	<0.001
		16	32	0.023	0.003	0.044	0.016
		16	64	0.038	0.018	0.059	<0.001
		16	128	0.047	0.027	0.067	<0.001
		32	128	0.024	0.003	0.044	0.013
Resampling[1]	T1ce	1	1.25	0.020	<0.001	0.039	0.047
		1	1.5	0.042	0.022	0.061	<0.001
		1	1.75	0.044	0.025	0.063	<0.001
		1	2	0.051	0.031	0.070	<0.001
		1.25	1.5	0.022	0.003	0.041	0.017
		1.25	1.75	0.024	0.005	0.044	0.006
		1.25	2	0.031	0.012	0.051	<0.001
	T2	1	1.5	0.024	0.006	0.043	0.003
		1	2	0.021	0.003	0.039	0.014
		1.25	1.5	0.021	0.003	0.039	0.016

[1]Performed with isotropic fashion. One dimension (mm) of the voxel is presented. CI, confidence interval; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted.

spective bin count (e.g., for T1ce, a bin count of 128 vs. a bin width of 6) in terms of the percentages of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility in all comparisons, with statistically significant distributional differences (Table 4).

Figures 5 and 6 for the T1ce sequence and Supplementary Figures S1 and S2 for the T2 sequence depict the reproducibility of radiomic features according to the feature classes and image types from which they were extracted. In the qualitative evaluation of these bar charts, there was no major trend deviation other than the original image against the general trend.

Discussion

In this study, we systematically investigated the influence of image preprocessing parameters (i.e., discretization and resampling) on the segmentation-based reproducibility of MRI radiomic features and found a significant impact. The bin width method yielded more reliable features than the bin count method. Using lower bin width values and higher resampling values produced more reproducible features.

Several studies have evaluated the influence of preprocessing and segmentation independently,[34] neglecting their influence on each other to a large extent. To our knowledge, very few studies have focused on the impact of preprocessing settings on segmentation-based reproducibility.[24,25] Additionally, no research has specifically assessed the impact of both image voxel resampling and gray-level discretization on the segmentation-based reproducibility of radiomic features.

Duron et al.[24] studied two independent MRI datasets of lachrymal gland tumors and breast lesions from two different centers, with two-dimensional delineations for each dataset. They evaluated six absolute (i.e., fixed bin width method) and eight relative

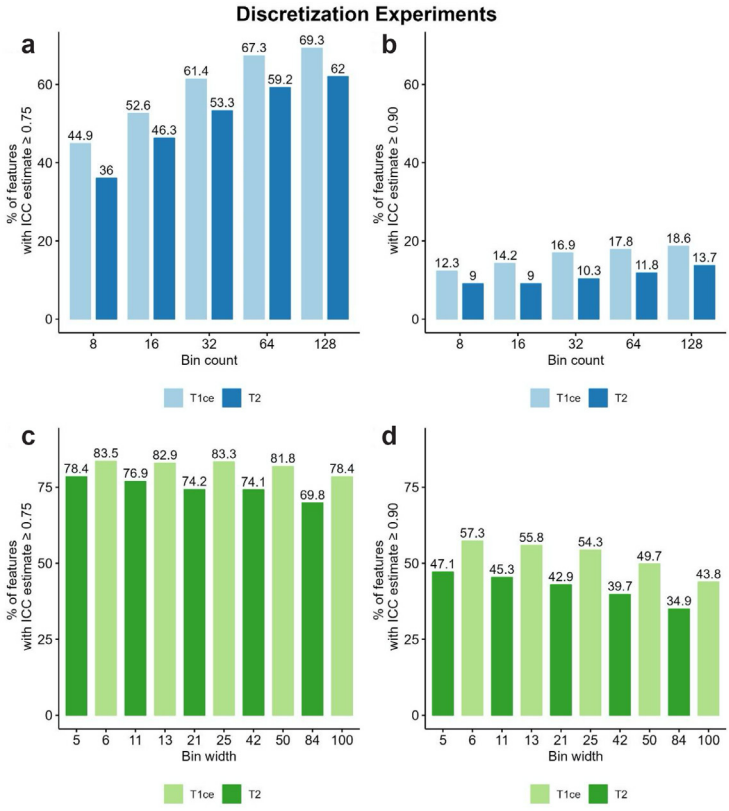


Figure 3. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility based on experiments with discretization parameters. Experiments with bin count **(a, b)** and bin width **(c, d)** on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. ICC, intraclass correlation coefficient.

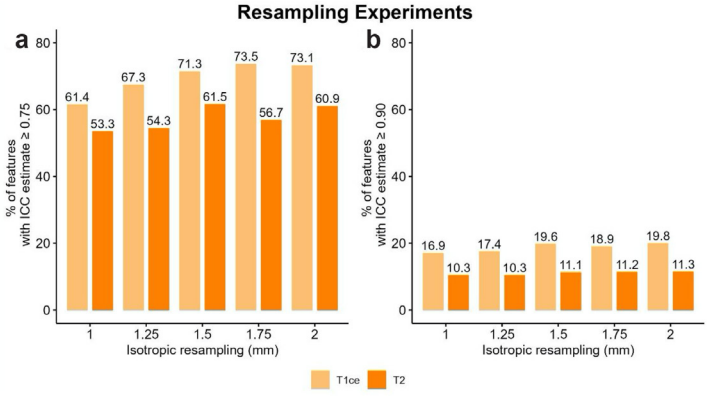


Figure 4. Percentage of reproducible features based on experiments on resampling parameters, using ICC cut-off values of 0.75 **(a)** and 0.90 **(b)** for good and excellent reproducibility, respectively. Experiments were performed on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. ICC, intraclass correlation coefficient.

(i.e., bin count method) discretization parameters and studied the distribution and highest number of replicable features for each technique. In addition, they utilized computer-generated delineations that were indicative of inter-observer variability. They observed that the discretization approach had a direct impact on feature repeatability, independent of observers, software, or method of delineation (simulated vs. human). Absolute discretization (i.e., the fixed bin width method) was recommended because it consistently produced statistically considerably more reproducible features than relative discretization. Large bin numbers or narrow bin widths produced the highest number of repeatable features in all experiments. They also underlined that, regardless of the selected method, detailed documentation is vital so that results can be replicated. Although the tumors and range of parameters were completely different in our study from those of Duron et al.[24], we observed similar trends in discretization experiments that confirmed and supported each other. Conversely, the most recent guidelines released by the IBSI,[5] and a recent seminal phantom study,[8] recommend relative discretization techniques (i.e., the bin count method) across disparate acquisitions. Despite the recommendations, some other studies have shown that the relative discretization method might not be the optimal technique.[24]

Lu et al.[25] investigated the robustness of PET/CT-based radiomic features in terms of segmentation and discretization and conducted experiments to study them in patients with nasopharyngeal carcinomas. In total, 50%–63% of their features had an ICC ≥0.8 for the segmentation experiments, whereas 21%–23% of features showed an ICC ≥0.8 for the discretization experiments. However, only 6 of 57 features (11%) had an ICC ≥0.8 for the simultaneous evaluation of both segmentation and discretization experiments. Although Lu et al.[25] used a methodology that was quite different from ours, their study was indeed successful in showing the impact of discretization on the segmentation-based reproducibility of the radiomic features.

Unlike the above-mentioned studies, we additionally experimented with resampling parameters and discovered that increasing resampling size resulted in improved segmentation-based reproducibility rates. This additional finding on resampling is contradictory to the studies on the phantom experiments regarding the reproducibility of radiomic feature values. For instance, in a very recent phantom study, Wichtmann et al.[8] recommended that resampled voxels should not be too far from the original voxel size regarding feature reproducibility.

Our experiments and previous studies indicate that both discretization and resampling parameters significantly impact the segmentation-based reproducibility of radiomic features, and the optimal parameters to achieve high reproducibility in feature values and segmentation-based reproducibility seem contradictory. For this reason, care should be taken to find the optimal parameters to achieve both feature value reproducibility and segmentation-wise reproducible features within the radiomic pipeline.

This study has several differences when compared with previous studies. First, the number of features was higher than that of previous studies and was as high as those in radiomics research publications that had a clinical purpose. Second, the analysis was not limited to discretization but included ex-

Table 4. Comparison of reproducible features between different discretization techniques

ICC cut-off	Sequence	Bin count vs. bin width	Statistic[1]	df	*P*[2]
0.75	T1ce	128 vs. 6	104.4	1	<0.001
		16 vs. 50	270.7	1	<0.001
		32 vs. 25	182.6	1	<0.001
		64 vs. 13	114.2	1	<0.001
		8 vs. 100	293.5	1	<0.001
	T2	128 vs. 5	120.5	1	<0.001
		16 vs. 42	242.0	1	<0.001
		32 vs. 21	152.5	1	<0.001
		64 vs. 11	124.3	1	<0.001
		8 vs. 84	307.8	1	<0.001
0.90	T1ce	128 vs. 6	398.1	1	<0.001
		16 vs. 50	365.0	1	<0.001
		32 vs. 25	380.7	1	<0.001
		64 vs. 13	386.7	1	<0.001
		8 vs. 100	325.4	1	<0.001
	T2	128 vs. 5	322.7	1	<0.001
		16 vs. 42	321.8	1	<0.001
		32 vs. 21	338.4	1	<0.001
		64 vs. 11	337.0	1	<0.001
		8 vs. 84	300.0	1	<0.001

[1]McNemar's chi-squared. [2]In all comparisons, bin width was superior to bin count in terms of proportions of reproducible features. ICC, intraclass correlation coefficient; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; df, degrees of freedom.

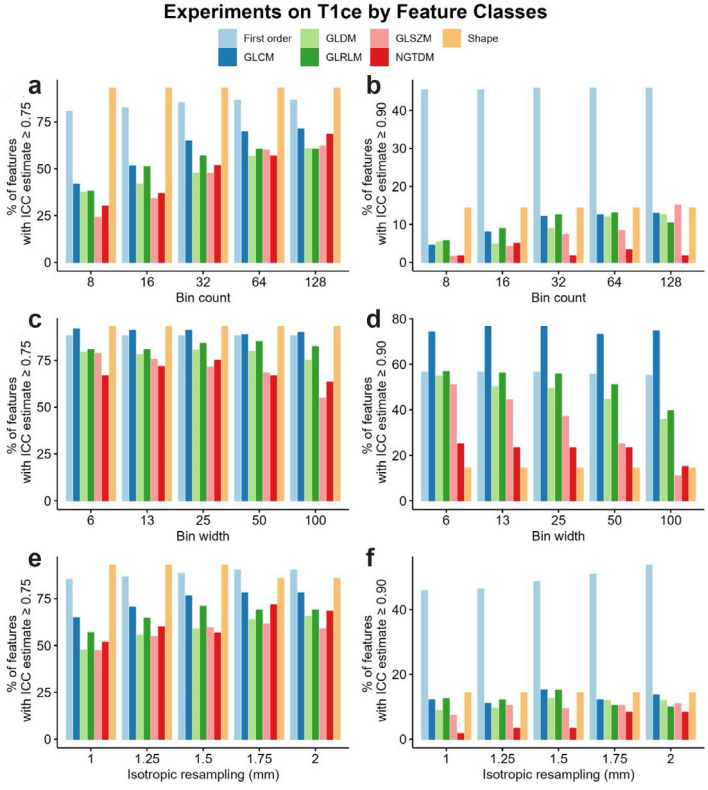


Figure 5. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by feature classes. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on contrast-enhanced T1-weighted (T1ce) sequence. ICC, intraclass correlation coefficient; GLCM, gray-level cooccurrence matrix; GLSZM, gray-level size zone matrix; GLRLM, gray-level run-length matrix; GLDM, gray-level dependence matrix; NGTDM, neighboring gray-tone difference matrix.

periments regarding resampling. These two preprocessing options commonly appear in open-source feature extraction software programs. Third, the experiments were performed in a different pathology (i.e., glioma), expanding the knowledge of the impact of preprocessing on segmentation-based reproducibility of radiomic features.

The public annotation dataset of BraTS 2021 was not used in the reproducibility experiments of this study because those data were based on a fusion of resultant annotations from several automated methods, first using the simultaneous truth and performance level estimation algorithm, followed by corrections applied by experts.[28] It would be difficult to perform and replicate the reproducibility experiments based on the public dataset, which may also not be representative of radiomics publications in general (not specifically those on gliomas) because those papers assessing segmentation reproducibility generally include at least two individual readers. For this reason, we segmented the dataset included in this study ourselves using the whole tumor volume to truly represent the segmentation-based reproducibility step of the radiomic studies.

Our experiments provided several practical points that might be considered in radiomic pipelines, associated publications, and clinical applications. First, image processing including discretization and voxel resampling has a considerable impact on the segmentation-based reproducibility of radiomic features; this should be considered as a means of improving the reproducibility of radiomic features that will be input to the following modeling stages. Second, the bin width method provided more reliable features than the bin count method in terms of segmentation-based reproducibility. Therefore, the bin width method should be favored in clinical studies. Third, using lower values for the bin width and higher values for the resampling provided more reproducible features. Given that there has been a lack of standardized preprocessing settings for discretization and resampling in the literature, these findings might provide guidance for end-users of the radiomic feature extraction tools. Fourth, due to their influence on the generation of reproducible inputs for modeling, our findings indicate that the preprocessing methods and their parameters must be defined in detail in published articles for radiomics models to be reliable.[35] According to a recent study, these essential radiomic parameters have been usually poorly reported in publications.[7] The recently published

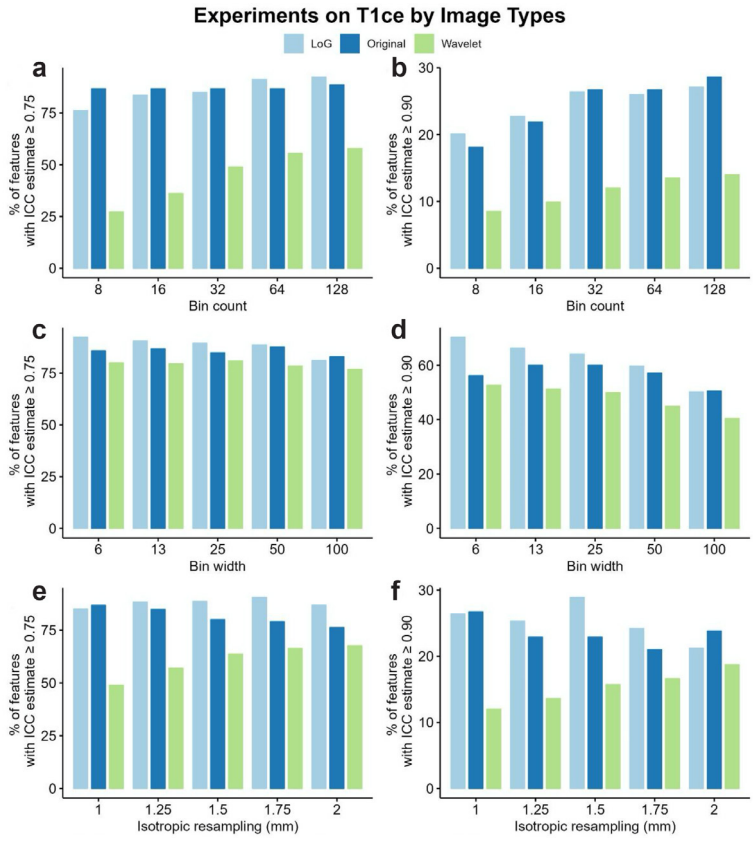


Figure 6. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by image types. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on contrast-enhanced T1-weighted (T1ce) sequence. ICC, intraclass correlation coefficient.

Checklist for Evaluation of Radiomics Research has also drawn attention to the same reporting issues.[9]

Our findings in this study should be interpreted with the following limitations.

First, the protocol for the acquisition of the BraTS 2021 challenge is not entirely clear. It is necessary to conduct research into the influence of the acquisition protocol (e.g., scanner type or acquisition settings) on image properties to gain a deeper comprehension of the behavior of radiomic features.

Second, our research was limited to a single imaging modality, two sequences, manual three-dimensional segmentation, a single tumor pathology, and gross tumor volume to remain manageable, considering the number of experiments conducted. However, we should acknowledge that every one of the aforementioned limitations may hamper the generalizability of the findings. We could also have added other alternatives to this study; however, that may have unnecessarily increased the complexity and workload, which was already high. This study aimed primarily to bring the attention of the radiomics community to the sensitivity of segmentation-based reproducibility to slight changes in two common preprocessing methods and offer reasonable settings. Alternative factors, such as different tumors, other MRI sequences, and different segmentation techniques, should be investigated as part of ongoing research.

Third, although significant and recommended by the IBSI guidelines,[5] the preprocessing techniques utilized in this study were only representative of a subset of the available options. However, the methods we used are available on the user interface of nearly all open-source radiomic feature extraction tools. The issue of standardization in radiomic studies may also involve scanner performance, acquisition protocols, acquisition sequence parameters, and data analysis techniques. However, we believe that the results of our study could be a step toward the standardization of radiomics.

Fourth, in our resampling experiments, the bin count was fixed. In light of the pairwise comparison experiments that were conducted with the final number of gray levels fixed, we anticipate observing a similar pattern when employing the bin width method. Additionally, when resampling images, we performed downsampling, as there has been no clear evidence on whether upsampling or downsampling methods are preferable.[2,5,8] However, although we considered the use of upsampling to be counterintuitive due to the addition of new voxels, it should be further explored in future experiments.

Fifth, the optimal settings for image processing to achieve the highest proportion of reproducible features were specific to the configuration used in this study. Our objective was not to identify absolute optimal values for all combinations of preprocessing settings. Consequently, no definitive conclusions should be drawn regarding the absolute best parameters (because, for example, they may be beyond the range of parameters used in the experiments) or the optimal sequence and discriminative performance.

Sixth, we did not test semi-automated or automated procedures in this study. Even with such techniques, a human touch or consensus segmentation is usually needed for correction, necessitating an analysis of feature reproducibility for segmentation, and supporting the need for conducting such a study.

In conclusion, to improve and standardize radiomic applications, every potential dependency of radiomic features on various parts of the radiomic workflow should be considered while developing a clinical or research project. In this study, the effect of image preprocessing parameters on the segmentation-based reproducibility of radiomic features from MRI was investigated systematically. Variations of image processing parameters related to discretization and resampling had a significant impact on the segmentation-based reproducibility of radiomic features within the scope of this study, regardless of MRI sequences. In terms of segmentation-based reproducibility, the bin width method yielded more reliable features than the bin count method. Using lower bin width values and higher resampling values produced more reproducible features. We recommend that these processing parameters be determined within the radiomic pipeline and transparently reported in radiomic publications. We anticipate that the implementation of our recommendations may facilitate the selection of more reproducible features and enhance the translation and generalizability of radiomics analyses. Considering the radiomics reproducibility crisis, extensive reproducibility studies are required before radiomics can be reliably implemented in routine clinical practice.

Conflict of interest disclosure

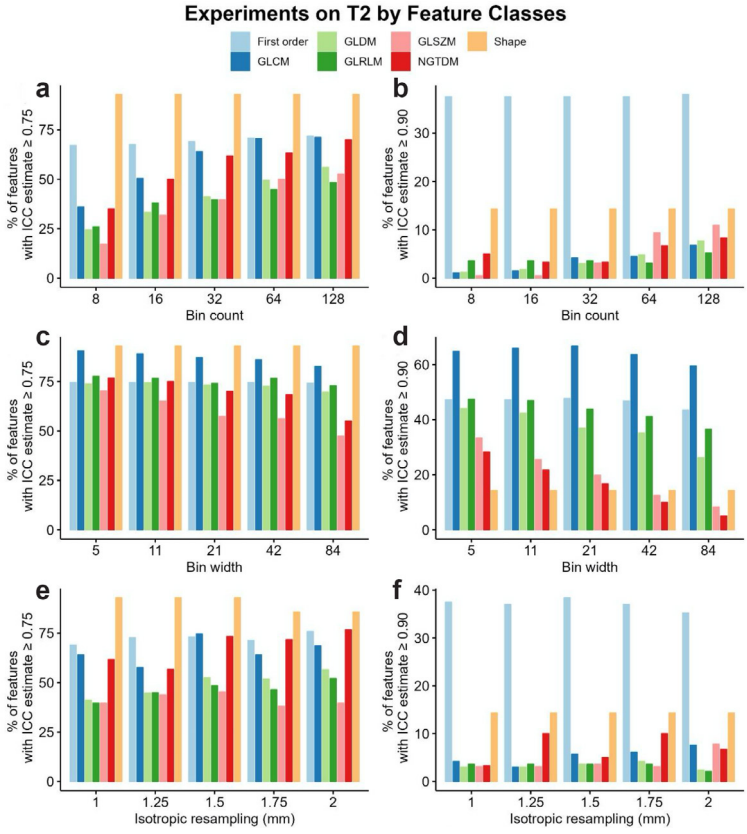
Burak Koçak, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

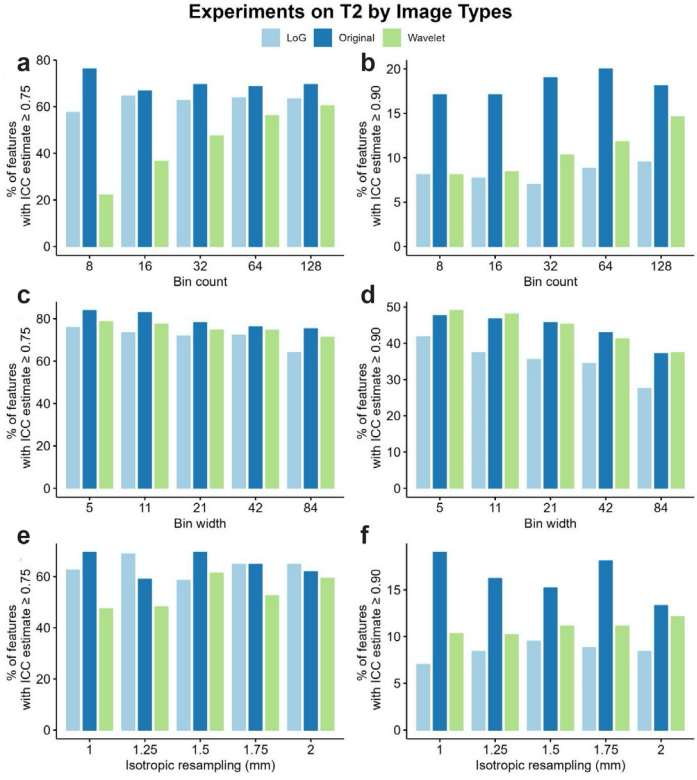
1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577. [CrossRef]
2. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91. [CrossRef]
3. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol*. 2020;93(1108):20190948. [CrossRef]
4. Kocak B, Baessler B, Cuocolo R, Mercaldo N, Pinto Dos Santos D. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *Eur Radiol*. 2023;33(11):7542-7555. [CrossRef]
5. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. [CrossRef]
6. Hagiwara A, Fujita S, Ohno Y, Aoki S. Variability and standardization of quantitative imaging: monoparametric to multiparametric quantification, radiomics, and artificial intelligence. *Invest Radiol*. 2020;55(9):601-616. [CrossRef]
7. Kocak B, Yuzkan S, Mutlu S, Bulut E, Kavukoglu I. Publications poorly report the essential RadiOmics ParametERs (PROPER): a meta-research on quality of reporting. *Eur J Radiol*. 2023;167:111088. [CrossRef]
8. Wichtmann BD, Harder FN, Weiss K, et al. Influence of image processing on radiomic features from magnetic resonance imaging. *Invest Radiol*. 2023;58(3):199-208. [CrossRef]
9. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [CrossRef]
10. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep*. 2019;9(1):4800. [CrossRef]
11. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography*. 2016;2(4):361-365. [CrossRef]
12. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1143-1158. [CrossRef]
13. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol*. 2019;19:33-38. [CrossRef]
14. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50(11):757-765. [CrossRef]
15. Gitto S, Cuocolo R, Emili I, et al. Effects of interobserver variability on 2D and 3D CT- and MRI-based texture feature reproducibility of cartilaginous bone tumors. *J Digit Imaging*. 2021;34(4):820-832. [CrossRef]
16. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. *Eur Radiol*. 2019;29(9):4765-4775. [CrossRef]
17. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428. [CrossRef]
18. Alis D, Yergin M, Asmakutlu O, Topel C, Karaarslan E. The influence of cardiac motion on radiomics features: radiomics features of non-enhanced CMR cine images greatly vary through the cardiac cycle. *Eur Radiol*. 2021;31(5):2706-2715. [CrossRef]
19. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur Radiol*. 2020;30(11):6241-6250. [CrossRef]
20. Ubaldi L, Saponaro S, Giuliano A, Talamonti C, Retico A. Deriving quantitative information from multiparametric MRI via radiomics: evaluation of the robustness and predictive value of radiomic features in the discrimination of low-grade versus high-grade gliomas with machine learning. *Phys Med*. 2023;107:102538. [CrossRef]
21. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124-1137. [CrossRef]
22. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. *Front Oncol*. 2021;11:633176. [CrossRef]
23. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol*. 2019;25(6):485-495. [CrossRef]
24. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One*. 2019;14(3):e0213459. [CrossRef]
25. Lu L, Lv W, Jiang J, et al. Robustness of radiomic features in [^{11}C]Choline and [^{18}F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. *Mol Imaging Biol*. 2016;18(6):935-945. [CrossRef]
26. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117. [CrossRef]
27. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024. [CrossRef]
28. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. Published online September 12, 2021. [CrossRef]
29. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798-819. [CrossRef]
30. Thakur S, Doshi J, Pati S, et al. Brain extraction on MRI scans in presence of diffuse glioma: multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *Neuroimage*. 2020;220:117081. [CrossRef]
31. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. [CrossRef]
32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. [CrossRef]
33. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15(2):155-163. [CrossRef]
34. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Phys Med*. 2019;61:44-51. [CrossRef]
35. Carré A, Klausner G, Edjlali M, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep*. 2020;10(1):12340. [CrossRef]

Supplementary Table S1. Patient identifiers

#	Identifier	#	Identifier
1	BraTS2021_00134	26	BraTS2021_01028
2	BraTS2021_00138	27	BraTS2021_01089
3	BraTS2021_00147	28	BraTS2021_01161
4	BraTS2021_00167	29	BraTS2021_01180
5	BraTS2021_00221	30	BraTS2021_01251
6	BraTS2021_00233	31	BraTS2021_01254
7	BraTS2021_00247	32	BraTS2021_01302
8	BraTS2021_00271	33	BraTS2021_01357
9	BraTS2021_00306	34	BraTS2021_01359
10	BraTS2021_00316	35	BraTS2021_01360
11	BraTS2021_00317	36	BraTS2021_01365
12	BraTS2021_00364	37	BraTS2021_01380
13	BraTS2021_00373	38	BraTS2021_01426
14	BraTS2021_00446	39	BraTS2021_01447
15	BraTS2021_00453	40	BraTS2021_01465
16	BraTS2021_00514	41	BraTS2021_01476
17	BraTS2021_00557	42	BraTS2021_01479
18	BraTS2021_00575	43	BraTS2021_01491
19	BraTS2021_00577	44	BraTS2021_01537
20	BraTS2021_00612	45	BraTS2021_01578
21	BraTS2021_00744	46	BraTS2021_01585
22	BraTS2021_00758	47	BraTS2021_01610
23	BraTS2021_00836	48	BraTS2021_01613
24	BraTS2021_01000	49	BraTS2021_01614
25	BraTS2021_01003	50	BraTS2021_01626



Supplementary Figures S1. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by radiomic feature classes. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on T2-weighted (T2) sequence. ICC, intraclass correlation coefficient; GLCM, gray-level cooccurrence matrix; GLSZM, gray-level size zone matrix; GLRLM, gray-level run-length matrix; GLDM, gray-level dependence matrix; NGTDM, neighboring gray-tone difference matrix.



Supplementary Figures S2. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by image types. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on T2-weighted (T2) sequence. ICC, intraclass correlation coefficient.