European Radiology
https://doi.org/10.1007/s00330-021-08282-1

MAGNETIC RESONANCE

Precision of MRI radiomics features in the liver and hepatocellular carcinoma

Guillermo Carbonell[1,2] · Paul Kennedy[1] · Octavia Bane[1] · Ammar Kirmani[1] · Maria El Homsi[3,4] · Daniel Stocker[1,5] · Daniela Said[1,6] · Pritam Mukherjee[7] · Olivier Gevaert[7] · Sara Lewis[1,3] · Stefanie Hectors[1] · Bachir Taouli[1,3]

Received: 16 March 2021 / Revised: 12 July 2021 / Accepted: 17 August 2021

Abstract

Objectives To assess the precision of MRI radiomics features in hepatocellular carcinoma (HCC) tumors and liver parenchyma.

Methods The study population consisted of 55 patients, including 16 with untreated HCCs, who underwent two repeat contrast-enhanced abdominal MRI exams within 1 month to evaluate: (1) test–retest repeatability using the same MRI system ($n = 28$, 10 HCCs); (2) inter-platform reproducibility between different MRI systems ($n = 27$, 6 HCCs); (3) inter-observer reproducibility ($n = 16$, 16 HCCs). Shape and 1st- and 2nd-order radiomics features were quantified on pre-contrast T1-weighted imaging (WI), T1WI portal venous phase (pvp), T2WI, and ADC (apparent diffusion coefficient), on liver regions of interest (ROIs) and HCC volumes of interest (VOIs). Precision was assessed by calculating intraclass correlation coefficient (ICC), concordance correlation coefficient (CCC), and coefficient of variation (CV).

Results There was moderate to excellent test–retest repeatability of shape and 1st- and 2nd-order features for all sequences in HCCs (ICC: 0.53–0.99; CV: 3–29%), and moderate to good test–retest repeatability of 1st- and 2nd-order features for T1WI sequences, and 2nd-order features for T2WI in the liver (ICC: 0.53–0.73; CV: 12–19%). There was poor inter-platform reproducibility for all features and sequences, except for shape and 1st-order features on T1WI in HCCs (CCC: 0.58–0.99; CV: 3–15%). Good to excellent inter-observer reproducibility was found for all features and sequences in HCCs (CCC: 0.80–0.99; CV: 4–15%) and moderate to good for liver (CCC: 0.45–0.86; CV: 6–25%).

Conclusions MRI radiomics features have acceptable repeatability in the liver and HCC when using the same MRI system and across readers but have low reproducibility across MR systems, except for shape and 1st-order features on T1WI. Data must be interpreted with caution when performing multiplatform radiomics studies.

Key Points

- *MRI radiomics features have acceptable repeatability when using the same MRI system but less reproducible when using different MRI platforms.*
- *MRI radiomics features extracted from T1 weighted-imaging show greater stability across exams than T2 weighted-imaging and ADC.*
- *Inter-observer reproducibility of MRI radiomics features was found to be good in HCC tumors and acceptable in liver parenchyma.*

Keywords Repeatability · Reproducibility · MRI radiomics · Liver · Hepatocellular carcinoma

✉ Bachir Taouli
bachir.taouli@mountsinai.org

[1] BioMedical Engineering and Imaging Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2] Department of Radiology, University Hospital Virgen de La Arrixaca, Murcia, Spain

[3] Department of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[5] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland

[6] Department of Radiology, Universidad de los Andes, Santiago, Chile

[7] Department of Radiology, Stanford University, Stanford, CA, USA

Published online: 26 September 2021

Abbreviations

ADC	Apparent diffusion coefficient
CCC	Concordance correlation coefficient
CV	Coefficient of variation
GLCM	Gray-level co-occurrence matrix
GLDM	Gray-level dependence matrix
GLRLM	Gray-level run length matrix
GLSZM	Gray-level size zone matrix
HCC	Hepatocellular carcinoma
IBSI	Image Biomarker Standardization Initiative
ICC	Intraclass correlation coefficient
NGTDM	Neighboring gray tone difference matrix
QIB	Quantitative imaging biomarker
QIBA	Quantitative Imaging Biomarkers Alliance
ROI	Region of interest
T1WIpre	T1-weighted imaging pre-contrast
T1WIpvp	T1-weighted imaging portal venous phase
T2WI	T2-weighted imaging
TE	Echo time
TR	Repetition time
VOI	Volume of interest

Introduction

The emerging field of radiomics, the extraction and analysis of large amounts of quantitative features from medical images, has gained popularity in the last decade [1, 2]. Previous studies have reported the ability of radiomics to characterize tumors and provide prognostication in different diseases [2–4]. Moreover, there is an aim to convert radiomic features into quantitative imaging biomarkers (QIBs), defined as objectively measured characteristics derived from an in vivo image as indicators of normal biological processes, pathogenic processes, or response to treatment [5, 6]. However, radiomics analyses consist of a complex workflow with several pre-processing steps [1, 7] that could drastically impact the result. In addition, variation in image acquisition parameters may affect generalizability of conclusions derived from radiomics analyses.

The Image Biomarker Standardization Initiative (IBSI) [8, 9], among other initiatives [10–12], has developed guidelines in an attempt to homogenize the radiomics process, and to create reproducible diagnostic and prognostic models; however, there is still lack of consensus on how to approach radiomics analyses. Therefore, precision studies assessing radiomics repeatability (evaluating features using identical or near-identical conditions) and reproducibility (evaluating features using different locations, operators, measuring systems, or other factors) are essential to determine the limitations of radiomics and identify the barriers for deployment in routine clinical workflows [5, 6].

There are several studies that have investigated the repeatability and reproducibility of radiomics features using different imaging modalities, especially on computed tomography (CT) [7, 13–15]. However, to date, only a few studies have evaluated magnetic resonance imaging (MRI) radiomics features repeatability, either in phantoms [16, 17] or in various cancers [18–21]. These studies found that repeatability of radiomics depends on multiple factors, such as the MRI sequence analyzed, or the precision test used to assess repeatability. Additionally, most of these studies used a single MRI sequence for radiomics extraction or were conducted under controlled acquisition and reconstruction parameters which is difficult to achieve in a regular clinical setting. There are also several studies analyzing the impact of acquisition and reconstruction parameters, and different pre-processing steps on MRI radiomics reproducibility, mainly on phantoms [16, 22–24], cervical cancer [18], and brain tumors [24–26]. However, to the best of our knowledge, there are no studies assessing MRI radiomics repeatability or reproducibility in liver parenchyma and/or hepatocellular carcinoma (HCC) tumors using multiple MRI sequences, which may be helpful for developing QIBs for liver disease and liver cancer characterization.

Thus, the aim of our study was to assess the precision of MRI radiomics features in liver and HCC tumors extracted from routine MRI sequences used in clinical protocols. This was tested in 3 different ways: (1) test–retest repeatability using the same MRI system; (2) inter-platform reproducibility using a combination of different MRI systems; and (3) inter-observer reproducibility from two different readers using the same MRI system and the same time point.

Methods

Patients

This single-center study, consisting of retrospective and prospective data analysis, was approved by our Institutional Review Board. Initially, 52 patients who underwent two consecutive abdominal MRI exams within 1 month between January 2017 and December 2018 were retrospectively included. The requirement of written informed consent was waived on this group. Patients were excluded due to incomplete MRI protocols ($n=8$), or severe imaging artifacts ($n=5$) resulting in a cohort of 39 patients that constituted *Group 1*. The reason for follow-up exams in this group was to rule out cholelithiasis/choledocholithiasis ($n=7$), follow-up of indeterminate liver or renal masses ($n=10$), and interval minimally invasive procedures or treatments (endoscopic retrograde cholangiopancreatography (ERCP) ($n=5$), biliary drain placement ($n=4$), nephrostomy placement ($n=1$), laparoscopic cholecystectomy ($n=3$), selective trans-arterial

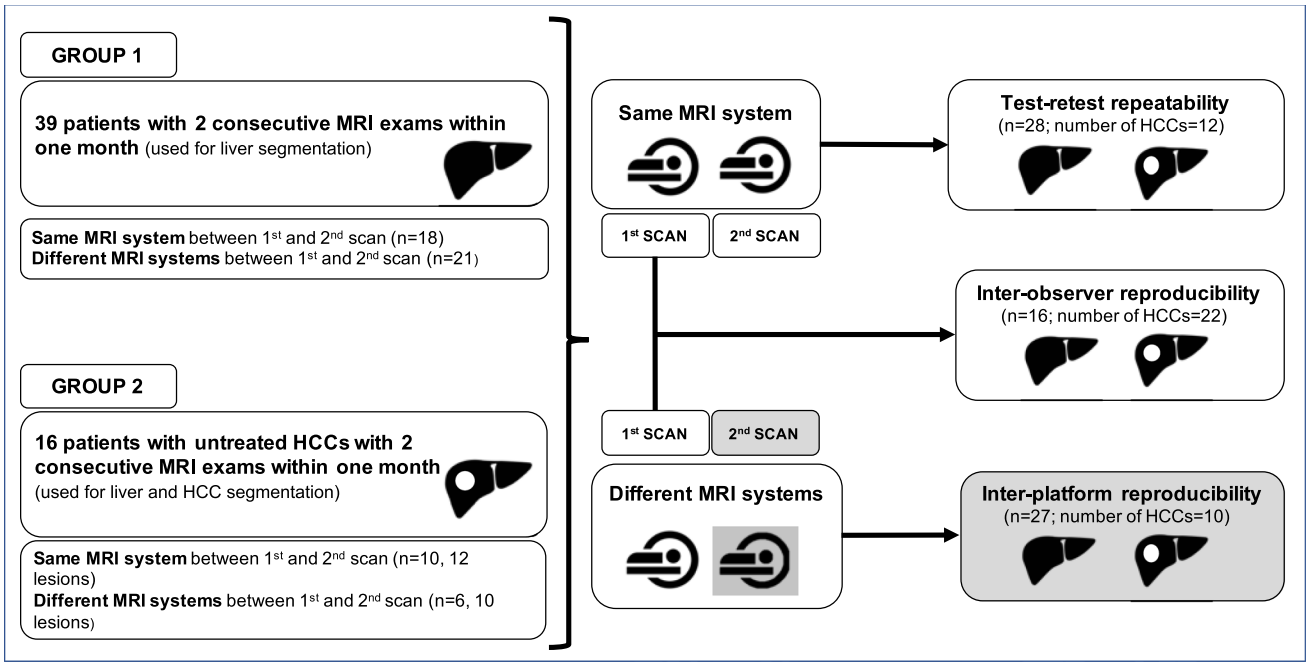

Fig. 1 Flow chart of the study design and the selection of the test–retest repeatability, inter-platform reproducibility, and inter-observer reproducibility groups

chemoembolization/radioembolization ($n=7$), microwave ablation therapy ($n=2$)). Patients in *Group 1* were only used for liver segmentation.

In addition, 16 prospectively recruited patients with untreated HCCs who underwent test–retest dynamic contrast-enhanced (DCE)-MRI examinations within 2 weeks between June 2013 and September 2014 constituted *Group 2*. Informed consent was obtained on this group. The DCE-MRI protocol was interrupted to include standard T1-weighted imaging (T1WI) acquisitions. For both groups, HCCs were diagnosed based on the Liver Reporting and Data System (LI-RADS) 2018 criteria [27]. Data from *Group 2* has been previously published [27–30]. The purpose of the previous studies was to evaluate multiparametric and quantitative MRI methods in HCC lesions. The assessment of repeatability/reproducibility of MRI radiomic features was beyond the scope of these studies. Patients in *Group 2* were used for liver and HCC segmentation.

The final study population consisted of 55 patients (combining *Groups 1* and *2*) and three defined tasks: (1) test–retest repeatability, (2) inter-platform reproducibility, (3) inter-observer reproducibility. There was overlap between the inter-observer reproducibility subset and the other two subsets, as all initial scans from patients with untreated HCC tumors (*Group 2*) were included for inter-observer assessment (Fig. 1). Patient characteristics are described in Table 1.

Table 1 Characteristics of patient population

Total patients	Test–retest repeatability ($n=28$)	Inter-platform reproducibility ($n=27$)	Inter-observer reproducibility ($n=16$)
Age (year)*	49 ± 18 (21–84)	58 ± 16 (23–81)	55 ± 13 (30–69)
Sex (male/female)	17/11	20/7	16/0
Time between scans (day)#	7 (1–28)	20 (3–30)	–
Number of patients with HCC	10 (35.7%)	6 (22.2%)	16 (100%)
Number of HCC lesions	12	10	22
Size of lesions (cm)*	5.6 ± 3.6 (1.7–11.3)	4.6 ± 3.3 (1.9–13.3)	5.1 ± 3.4 (1.7–13.3)

*Mean ± standard deviation (range)

#Mean (range)

MRI protocol

The test–retest repeatability experiment was performed using the same MRI system, and the inter-platform reproducibility task was performed using MRI systems from two different vendors. For the inter-observer reproducibility task, we used the first scan of the group of patients with untreated HCC (*Group 2*) for both test–retest repeatability and inter-platform reproducibility tasks. Description of MR acquisition parameters and MRI systems can be found in Table 2 and Electronic Supplementary Material 1 (ESM1).

Standard imaging protocols were used to obtain T1WI pre-contrast (T1WIpre) and during portal venous phase after contrast injection (T1WIpvp), T2WI using single-shot fast spin echo (SSFSE), and diffusion-weighted imaging (DWI). Other important post-contrast sequences for HCC assessment, such as arterial phase, were not used in this study as the subset of patients with untreated HCCs underwent DCE-MRI which did not include a clinical sequence during the arterial phase. The contrast agent used between first and second scan was the same in all patients, including gadobutrol (Gadavist/Gadovist®, Bayer Healthcare, $n = 23$), gadobenate dimeglumine (MultiHance®, Bracco, $n = 17$), or gadoxetate disodium (Eovist®, Bayer Healthcare, $n = 15$).

Image processing

Liver and tumor segmentation

One circular region of interest (ROI) measuring 30 mm in diameter was manually placed by a single observer (G.C., 5 years of abdominal MRI experience) within the liver parenchyma on T1WIpre, T1WIpvp, T2WI, and ADC from the first and second MRI studies. ROIs were located within the right hepatic lobe avoiding the capsule, large hepatic vessels, tumors, and treated areas. Afterwards, a volume of interest (VOI) was manually delineated by the same reader to segment the entire volume of each HCC on the subset of patients with untreated lesions (*Group 2)*. All VOIs were drawn on all slices where the lesion was visible. To minimize intra-observer variability between first and second MRI scan segmentations in the liver, images from the first and second scan for each patient were loaded at the same time on the software and anatomical landmarks were used to

Table 2 MRI acquisition parameters (mean, range in parentheses)

		MRI SYSTEMS							
	Acquisition parameters	GE Optima MR450w	GE Signa HDxt	GE Discovery MR750*	Siemens Aera	Siemens Avanto	Siemens Amira	Siemens Biograph	Siemens Skyra
T1WI	TR (ms)	4.2 (3.5–5.1)	3.4 (2.8–4.2)	2.9	4.6 (3.4–4.9)	3.5 (3.3–3.8)	4.1 (3.1–4.5)	3.6 (3.5–3.6)	3.1 (2.8–3.2)
	TE (ms)	1.3 (1.2–1.4)	1.5 (1.1–1.8)	1.4	2.2 (1.6–2.4)	1.3 (1.2–1.4)	1.8 (1.5–2.1)	1.6 (1.5–1.6)	1.4 (1.1–1.6)
	Pixel size (mm)	0.8 (0.7–0.9)	1.4 (1.2–1.6)	0.9	1.3 (1.2–1.6)	0.7 (0.6–0.7)	1.4 (1.2–1.7)	1.6 (1.5–1.6)	1.3 (1.2–1.6)
	Slice thickness (mm)	4.6 (4.6–5.0)	3.8 (3.1–5.0)	4.8	2.9 (2.5–3.5)	3.5 (3.0–4.2)	3.3 (2.5–5.0)	3.0 (3.0–3.0)	3.0 (2.0–5.0)
	Matrix	224×160–256×160	256×151–288×173	320×160	256×125–288×198	256×125–256×130	256×151–320×203	256×125–256×146	288×213–256×160
T2WI	TR (ms)	1080 (501–2800)	902 (508–1300)	1200	1169 (800–1300)	900 (900–900)	1205 (570–1300)	1300 (1200–1400)	1178 (569–1300)
	TE (ms)	236 (91–242)	229 (218–239)	181	100 (91–203)	238 (238–238)	190 (90–239)	91 (91–91)	99.9 (91–239)
	Pixel size (mm)	1.5 (1.3–1.8)	1.4 (1.3–1.6)	0.9	1.5 (1.3–1.7)	1.3 (1.3–1.4)	1.5 (1.2–1.7)	1.5 (1.5–1.5)	1.5 (1.4–1.6)
	Slice thickness (mm)	6.9 (5.0–7.0)	7.1 (7.0–7.2)	6.0	7.2 (7.2–8.4)	6.4 (6.0–7.2)	7.2 (7.0–7.4)	7.2 (7.2–7.2)	7.2 (7.0–7.2)
	Matrix	256×192	256×192–256×198	320×192	256×144–256×205	256×192–256×205	256×192–256×213	256×167–256×198	256×144–256×243
DWI	TR (ms)	14,367 (3500–20,000)	3867 (3600–4200)	3000	3997 (2500–4500)	6867 (3600–9000)	4431 (3600–7400)	9279 (6330–11,910)	4527 (3600–7450)
	TE (ms)	61 (59–63)	72 (67–78)	55	79 (75–80)	80 (77–82)	79 (68–89)	72 (70–73)	73 (66–74)
	Pixel size (mm)	1.5 (1.3–1.8)	1.3 (1.0–1.6)	1.7	1.3 (1.1–2.3)	1.0 (0.9–1.2)	1.2 (0.9–1.5)	1.3 (1.3–1.4)	1.2 (1.1–1.6)
	Slice thickness (mm)	6.1 (6.0–7.0)	7.5 (7.0–8.4)	8	7.1 (6.0–8.4)	7.5 (7.0–8.4)	7.5 (6.0–8.4)	7.0 (7.0–7.0)	7.3 (7.0–8.4)
	Matrix	128×128–144×128	128×80–160×128	160×128	160×80–160×132	160×128–192×168	128×80–192–168	128×128–160×102	128×80–160×160

T1WI T1-weighted imaging, *T2WI* T2-weighted imaging, *DWI* diffusion-weighted imaging, *TR* repetition time, *TE* echo time. GE Optima MRW450, Siemens Aera, or Siemens Skyra were used in the repeatability cohort for scan and re-scan; and a combination between GE Optima MRW450, GE Signa HDxT, GE Discovery MR750, Siemens Aera, Siemens Avanto, Siemens Amira, Siemens Biograph, and Siemens Skyra for scan and re-scan was used on the inter-platform cohort

keep the ROI location as similar as possible between scans. To assess inter-observer reproducibility, a second observer (M.E.H., 2 years of abdominal MRI experience) placed ROIs and VOIs in the liver parenchyma and HCCs, respectively, on each sequence of the first MRI study in the subset of patients with untreated HCCs. The second observer was blinded to the first observer segmentations. Examples of VOI and ROI placement are shown in Fig. 2 and ESM 2. All ROIs and VOIs were prescribed using software compliant with the IBSI guidelines (Olea sphere® 3.0, Olea Medical).

Image pre-processing and feature extraction

Spatial resampling was performed using nearest neighbor interpolation to create isotropic voxels ($1.0 \times 1.0 \times 1.0 mm^3$) to allow comparison between image data [9]. Signal intensity normalization was performed on T1WIpre, T1WIpvp, and T2WI sequences as previously described [17]. No normalization approach was performed on ADC as it reflects a quantitative property. A 64-fixed bin number was used for intensity discretization, as recommended by IBSI guidelines [9].

After performing the pre-processing steps, original intensity-based histogram or 1st-order features ($n = 19$) and original texture or 2nd-order features ($n = 73$) including gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), and neighboring gray tone difference matrix (NGTDM) were extracted from each ROI placed on the liver; and original shape features ($n = 16$), original 1st-order features ($n = 19$), and original 2nd-order features ($n = 73$) were extracted from VOIs delineated on HCCs. Shape-based features were extracted from VOIs placed on HCC tumors but not from liver ROIs because most of these features are dependent on the three-dimensional surfaces [28].

Statistical analysis

According to QIBA, specific tests should be used to assess the precision of QIBs for different scenarios [5]. For test–retest repeatability, a measurement of precision that occurs with identical/near identical conditions, we used the intraclass correlation coefficient (ICC). For inter-platform and inter-observer reproducibility assessment, where the measuring system or the readers are different, respectively, we used the concordance correlation coefficient (CCC). Furthermore, as assessing repeatability and reproducibility by calculation of ICC and CCC might not be sufficient since these calculations are known to depend on the natural variance of the underlying data [7], we have also calculated the coefficient of variation (CV), another precision test commonly used for repeatability and reproducibility experiments based on intra-subject variability. Detailed analysis of ICC, CCC, and CV calculation methodology can be found in ESM 3.

ICC for reporting test–retest repeatability and CCC for reporting inter-platform and inter-observer reproducibility were classified as follows: excellent (ICC/CCC ≥ 0.9); good (0.75–0.89); moderate (0.5–0.75); or poor (< 0.5) [18]. Regarding repeatability/reproducibility analysis using CV, the classification was as follows: excellent (CV $\leq 10\%$); good (11–20%), moderate (21–30%), and poor ($> 30\%$) [29, 30]. Results for each group of radiomic features (shape and 1st and 2nd order) within liver ROIs and VOIs delineated on HCCs for each MRI sequence were reported as median ICC/CCC with interquartile ranges, and as median CV. All analyses were performed using MatLab.

Results

Patients

The three tasks for assessing radiomics precision were defined as follows: (1) test–retest repeatability—$n = 28$

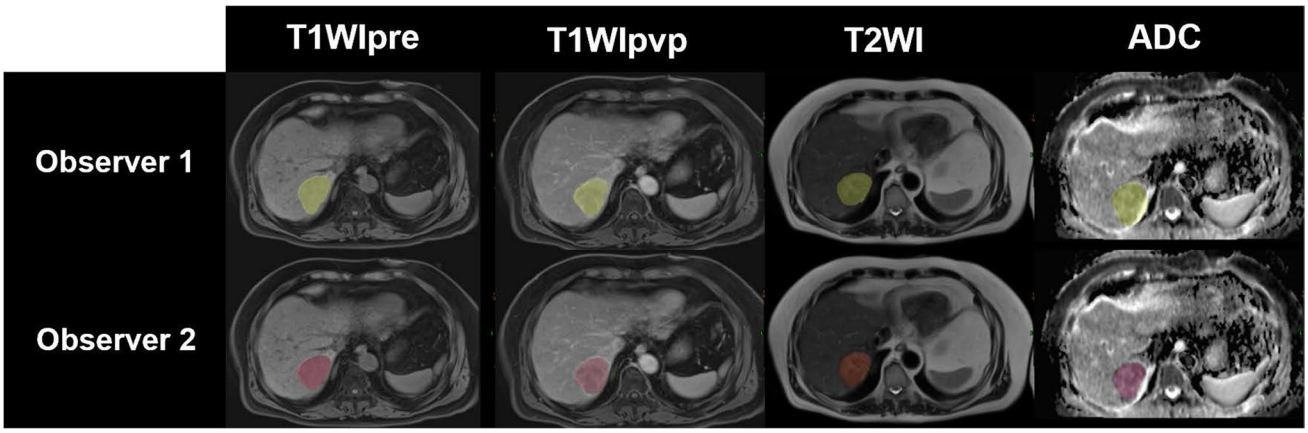

Fig. 2 Volume of interest delineation of HCC tumors performed by two different observers in T1WIpre, T1WIpvp, T2WI, and ADC

patients (*Group 1*, $n = 18$; *Group 2*, $n = 10$) with two consecutive abdominal MRI studies using the same MRI system; (2) inter-platform reproducibility—$n = 27$ patients (*Group 1*, $n = 21$; *Group 2*, $n = 6$) with two consecutive abdominal MRIs using different MRI systems; and (3) inter-observer reproducibility—$n = 16$ patients (all from *Group 2*) with one abdominal MRI on the same MRI system.

Test–retest repeatability

HCC VOIs: test–retest repeatability was good to excellent (ICC ≥ 0.75 and CV ≤ 20%) for shape features on T1WIpre, T1WIpvp, T2WI, and ADC, and for 1st- and 2nd-order features on T1WIpvp; and moderate (ICC 0.5–0.75; and CV ≤ 30%) for 1st- and 2nd-order features on T1WIpre, T2WI, and ADC.

Liver ROIs: test–retest repeatability was moderate (ICC 0.5–0.75; and CV ≤ 30%) for 1st- and 2nd-order features on T1WIpre and T1WIpvp, and for 2nd-order features on T2WI. Repeatability was poor (ICC < 0.5; or CV > 30%) for 1st-order features on T2WI, and for 1st- and 2nd-order features on ADC.

Detailed test–retest repeatability results are shown in Table 3 and Fig. 3.

Inter-platform reproducibility

HCC VOIs: inter-platform reproducibility was good to excellent (CCC ≥ 0.75; and CV ≤ 20%) for shape features on T1WIpre and T1WIpvp, and for 1st-order features on T1WIpvp; and moderate (CCC 0.5–0.75; and CV ≤ 30%) for 1st-order features on T1WIpre. Reproducibility was poor (CCC < 0.5; or CV > 30%) for 2nd-order features on T1WIpre and T1WIpvp, and for shape and 1st- and 2nd-order features on T2WI and ADC.

Liver ROIs: inter-platform reproducibility was poor (CCC < 0.5; or CV > 30%) for 1st- and 2nd-order features on T1WIpre, T1WIpvp, T2WI, and ADC.

Table 3 Overall repeatability and reproducibility of radiomics features per group and sequence

	T1WIpre		T1WIpvp		T2WI		ADC	
Test–retest repeatability	ICC (IQR)	CV	ICC (IQR)	CV	ICC (IQR)	CV	ICC (IQR)	CV
HCC VOI								
- Shape	0.99 (0.94–0.99)	4%	0.99 (0.79–0.99)	3%	0.99 (0.75–0.99)	4%	0.98 (0.40–0.99)	13%
- 1st order	0.66 (0.54–0.83)	14%	0.76 (0.68–0.82)	11%	0.95 (0.85–0.97)	23%	0.58 (0.30–0.82)	24%
- 2nd order	0.64 (0.54–0.84)	18%	0.75 (0.55–0.85)	13%	0.70 (0.46–0.84)	19%	0.53 (0.20–0.68)	29%
Liver ROI								
- 1st order	0.53 (0.50–0.60)	18%	0.56 (0.29–0.73)	18%	0.77 (0.58–0.89)	40%	0.34 (0.05–0.72)	17%
- 2nd order	0.73 (0.46–0.87)	12%	0.54 (0.33–0.73)	15%	0.53 (0.44–0.62)	19%	0.18 (−0.04–0.39)	21%
Inter-platform reproducibility	CCC (IQR)	CV	CCC (IQR)	CV	CCC (IQR)	CV	CCC (IQR)	CV
HCC VOI								
- Shape	0.99 (0.94–0.99)	3%	0.99 (0.87–0.99)	6%	0.42 (0.22–0.90)	14%	0.11 (−0.03–0.82)	19%
- 1st order	0.58 (0.50–0.79)	15%	0.76 (0.65–0.92)	11%	−0.06 (−0.16–0.43)	65%	0.16 (−0.75–0.62)	26%
- 2nd order	0.48 (0.31–0.66)	22%	0.49 (0.07–0.66)	25%	0.43 (−0.13–0.66)	27%	0.33 (0.00–0.52)	32%
Liver ROI								
- 1st order	0.20 (0.14–0.25)	27%	0.10 (0.04–0.16)	26%	0.35 (0.13–0.58)	32%	0.10 (0.00–0.28)	62%
- 2nd order	0.19 (0.13–0.31)	23%	0.28 (0.15–0.43)	22%	0.27 (0.17–0.41)	17%	0.02 (−0.13–0.06)	36%
Inter-observer reproducibility	CCC (IQR)	CV	CCC (IQR)	CV	CCC (IQR)	CV	CCC (IQR)	CV
HCC VOI								
- Shape	0.99 (0.71–0.99)	4%	0.99 (0.78–0.99)	4%	0.99 (0.90–0.99)	4%	0.99 (0.96–0.99)	4%
- 1st order	0.95 (0.93–0.99)	6%	0.97 (0.94–0.98)	5%	0.99 (0.93–0.99)	8%	0.96 (0.93–0.98)	8%
- 2nd order	0.95 (0.90–0.98)	8%	0.97 (0.94–0.98)	7%	0.80 (0.63–0.93)	15%	0.95 (0.90–0.98)	11%
Liver ROI								
- 1st order	0.65 (0.39–0.88)	10%	0.86 (0.78–0.94)	6%	0.73 (0.47–0.97)	25%	0.75 (0.71–0.79)	9%
- 2nd order	0.61 (0.34–0.77)	12%	0.61 (0.28–0.81)	13%	0.45 (0.38–0.52)	15%	0.78 (0.60–0.87)	10%

Intraclass correlation coefficients (ICC) and concordance correlation coefficients (CCC) are represented as means with interquartile ranges between brackets. Coefficients of variation (CV) are represented as the mean percentage

T1WIpre T1-weighted imaging pre-contrast, *T1WIpvp* T1-weighted imaging portal venous phase, *T2WI* T2-weighted-imaging, *ADC* apparent diffusion coefficient, *Liver ROI* liver region of interest, *HCC VOI* hepatocellular carcinoma volume of interest

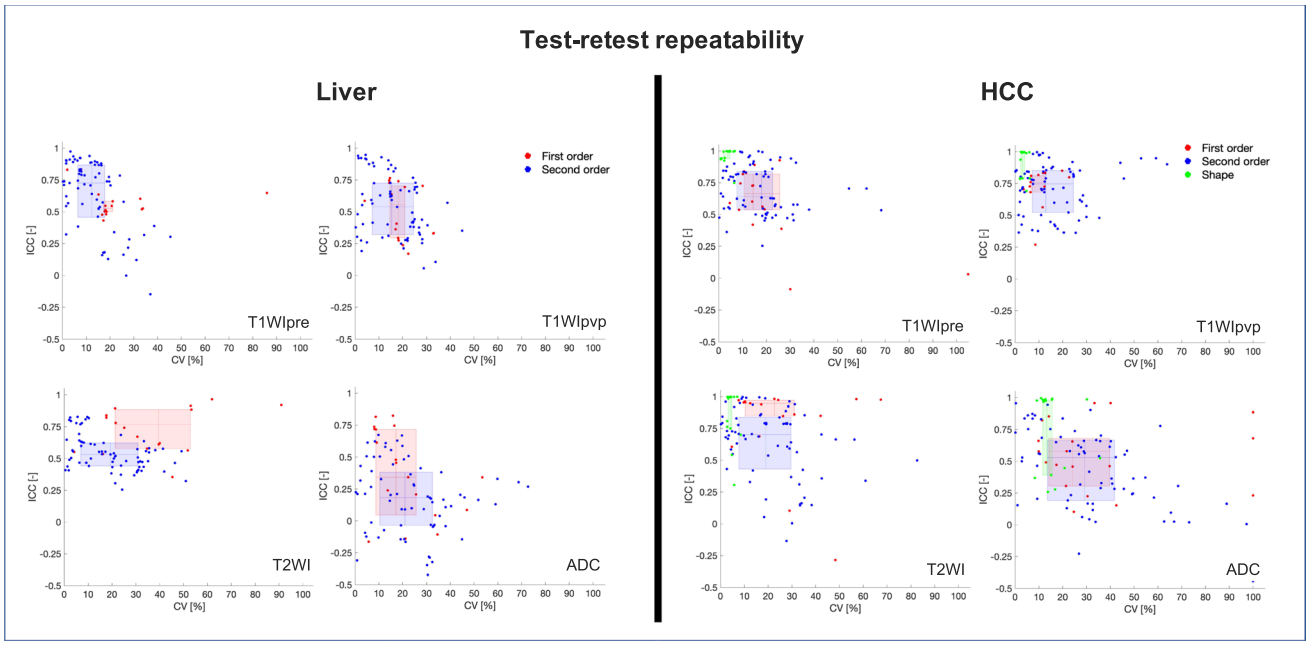

Fig. 3 Test–retest repeatability of radiomics features in HCC tumors and liver parenchyma illustrated by ICC and CV (medians, lines; IQR, boxes) per group of features (red dots: 1st-order features; blue dots: 2nd-order features; green dots: shape features) on T1WIpre, T1WIpvp, T2WI, and ADC

Detailed inter-platform reproducibility results are shown in Table 3 and Fig. 4.

Inter-observer reproducibility

HCC VOIs: inter-observer reproducibility was good to excellent (CCC ≥ 0.75; and CV ≤ 20%) for shape and 1st- and 2nd-order features on T1WIpre, T1WIpvp, T2WI, and ADC.

Liver ROIs: inter-observer reproducibility was good to excellent (CCC ≥ 0.75; and CV ≤ 20%) for 1st-order features on T1WIpvp, and 1st- and 2nd-order features on ADC; moderate (CCC 0.5–0.75; and CV ≤ 30%) for 1st- and 2nd-order features on T1WIpre, 2nd-order features on T1WIpvp, and 1st-order features on T2WI; and poor (CCC < 0.5; or CV > 30%) for 2nd-order features on T2WI.

Detailed inter-observer reproducibility results are shown on Table 3 and Fig. 5. Further analysis of radiomics features repeatability and reproducibility, including the number and percentage of most robust features for liver ROIs and HCC VOIs per MRI sequence, are available in Electronic Supplementary Material (ESM 4–14).

Discussion

The key findings of our study are as follows: (1) MRI radiomics features in HCC and liver parenchyma show relative stability when using the same MRI system; (2) MRI radiomics features exhibit a substantial drop in reproducibility on the inter-platform cohort on all sequences, with T1WI sequences being more stable than T2WI and DWI; and (3) MRI radiomics show excellent inter-observer reproducibility in HCC and moderate to good inter-observer reproducibility in liver parenchyma.

MRI radiomics quantification may represent a promising tool for patient management, and for selection for targeted therapies in HCC. It has been applied to predict tumor histopathology [31], immuno-oncologic characteristics [32, 33], tumor response, and patient outcome [34] in patients with HCC. Thus, precision studies assessing radiomics repeatability and reproducibility are key to implement this kind of analysis in the clinical practice.

Compared to previous studies, we found slightly lower repeatability for T1WI and T2WI sequences but these findings were expected as one of these studies was performed under more controlled conditions on an MRI phantom using T1WI, T2WI, and FLAIR sequences [16] and in patients with cervical cancer using only T2WI sequences [18].

Results from our inter-platform task showed a substantial drop in radiomics features reproducibility for 1st- and 2nd-order feature groups across all MR sequences, with a less pronounced decrease on T1WI sequences. This drop in radiomics reproducibility is likely related to acquisition parameters, reconstruction, and field strength variation between MRI systems from the same or different vendors. There are a few studies describing a similar impact on

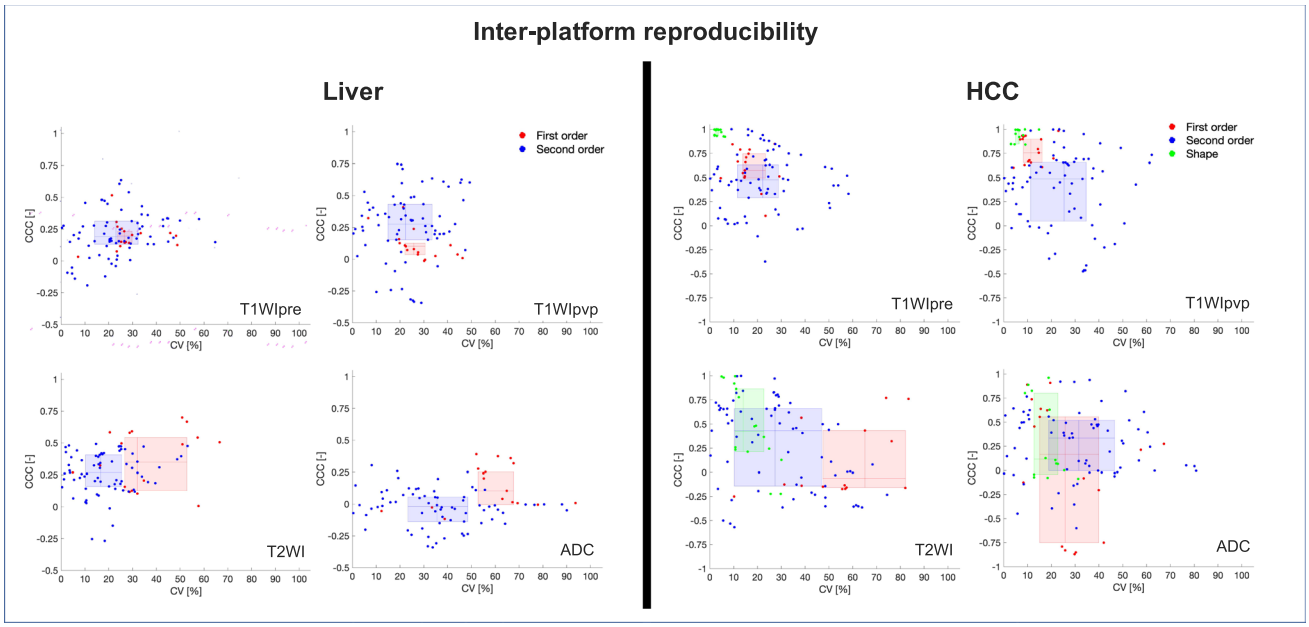

Fig. 4 Inter-platform reproducibility of radiomics features in HCC tumors and liver parenchyma illustrated by CCC and CV (medians, lines; IQR, boxes) per group of features (red dots: 1st-order features; blue dots: 2nd-order features; green dots: shape features) on T1WIpre, T1WIpvp, T2WI, and ADC

radiomics reproducibility when using different TR, TE, and voxel size in phantoms [22] or different MRI scanners between test and re-test in patients with cervical cancer [18] and glioblastoma [25].

In our study, a routine MRI follow-up scan using the same MRI system invariably incurred minor acquisition parameter variations between scans; however, these variations were amplified when the follow-up scan was performed on another

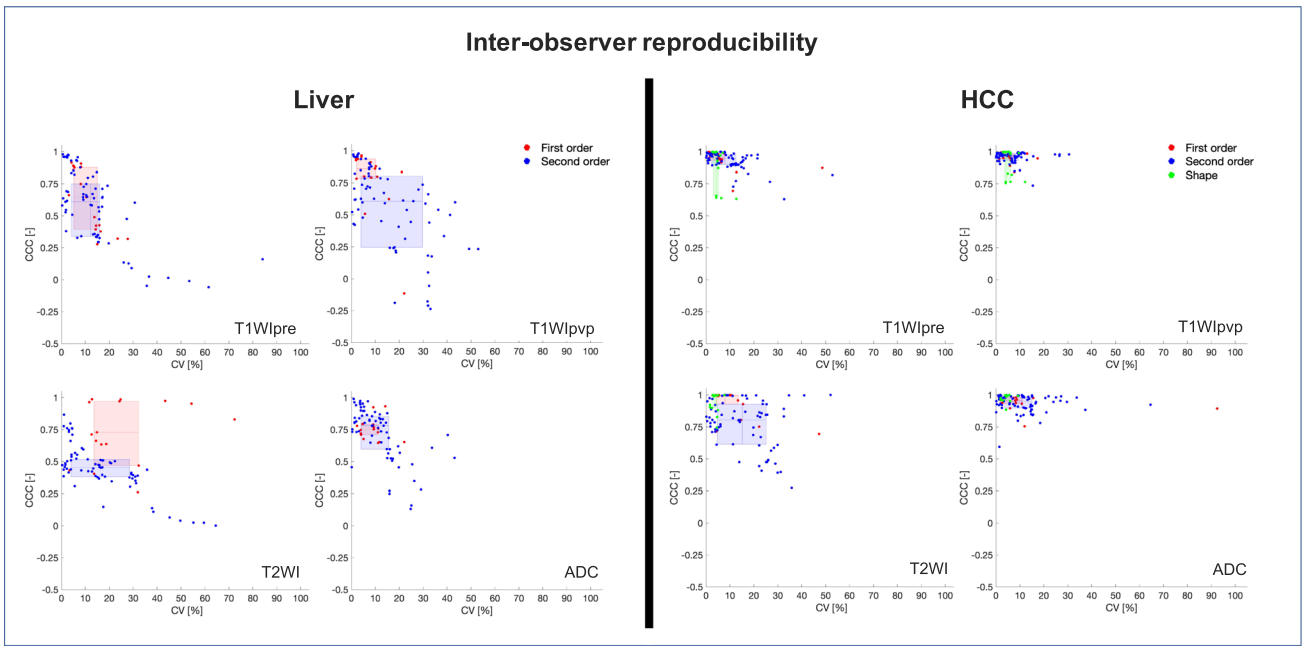

Fig. 5 Inter-observer reproducibility of radiomics features in HCC tumors and liver parenchyma illustrated by CCC and CV (medians, lines; IQR, boxes) per group of features (red dots: 1st-order features; blue dots: 2nd-order features; green dots: shape features) on T1WIpre, T1WIpvp, T2WI, and ADC

MR system. This should be considered when designing multi-center and retrospective studies using different MRI platforms as it may affect the outcome of radiomics analyses. Additionally, we cannot rule out that structural, physiological, and pathological changes within liver parenchyma and tumors between the first and second scans may cause a drop of feature repeatability and reproducibility as we are using routine MRI scans to assess radiomics precision. However, we assume that those factors would have minimum impact as the follow-up MRI scans were performed within a short period of time (less than 1 month).

Additionally, our analysis highlighted differences in feature stability between MRI sequences. Overall, T1WI sequences showed higher feature repeatability and reproducibility than T2WI and ADC acquisitions on test–retest and inter-platform subsets, respectively. Some authors have shown that MRI features extracted from high-resolution sequences in phantom studies are more stable than those from low-resolution sequences, even after performing spatial resampling of the images [16, 35]. T1WI sequences used in our study consisted of volumetric sequences, with smaller pixel size, pixel spacing, and slice thickness than T2WI and ADC, which may potentially explain their higher stability in an inter-platform setting. On the other hand, the lower feature repeatability on ADC could be explained by the lower spatial resolution of DWI and the fact that DWI is acquired free-breathing, including inputs from several *b*-values, which may be affected by motion. There are some discrepancies on the literature on this topic, with authors identifying good ADC features stability, with 25–29% stable features across different tissues, and different MR systems and vendors [20] while others state that results could vary dramatically depending on the processing configuration [19]. These differences could also be explained by the different pre-processing steps used in this study compared to previous work. There remains a lack of consensus on how to approach these pre-processing steps for different sequences. In future studies, different pre-processing steps should be evaluated to improve feature stability across MRI sequences.

Knowledge of inter-observer reproducibility is important before incorporating radiomics analysis in a routine clinical setting or in clinical trials. We found moderate to good inter-observer reproducibility for ROIs within the liver and excellent reproducibility for VOIs placed on HCCs across all groups of radiomic features and sequences in our cohort. Single-slice ROI placement in liver parenchyma is more susceptible to inter-observer variability compared to tumor delineation due to the absence of specific anatomical landmarks which help guide ROI placement. Furthermore, volumetric segmentation usually encompasses a larger sample of tissue which may improve radiomics stability. We chose this methodology as it reflects the current practice for quantitative liver assessment which leverages the general homogeneity of liver parenchyma throughout the organ; e.g., iron overload evaluation, fat-fraction quantification, and liver stiffness measurement using MR elastography are determined over a limited area. In contrast, HCC heterogeneity necessitates volumetric assessment to avoid sampling bias. Our results match with previous MRI phantom [36], and cervical cancer [18] studies. Conversely, a study carried out by Saha et al [37] on breast cancer MRI exams showed inter-observer variability assessed in breast tumors was higher than that in normal fibro-glandular tissue. Currently, manual segmentation performed by experienced readers is considered the standard of reference, but it is a time-consuming task and may be affected by inter-observer variability as we show. Thus, precise fully automated segmentation tools should be implemented to reduce analysis time and minimize operator interaction [1, 7].

Lastly, shape features extracted from HCC VOIs showed excellent repeatability on all sequences (ICCs between 0.98 and 0.99), and excellent inter-platform reproducibility on T1WI sequences (CCCs between 0.99 and 0.99), with poor reproducibility for T2WI and ADC. These findings were expected and widely concordant with several studies, both on CT [38, 39] and MRI [18].

The main limitations of our study were that we used retrospective data from a single center. We used a single pre-processing setting for all sequences, except for ADC where we did not perform a normalization approach. However, there is no consensus concerning the optimal pre-processing steps for each sequence and group of features. Further research on this topic is desirable. Additionally, we did not assess the exact influence of different MRI acquisition parameters and different field strengths in feature repeatability and reproducibility. We also assessed only one post-contrast sequence to evaluate feature variability. In the future, we will evaluate multiple post-contrast sequences to assess the impact of contrast administration on radiomics robustness.

In conclusion, our results show acceptable repeatability of MRI radiomics features of HCC tumors and liver parenchyma when using the same MRI system, with a drop in reproducibility when performing studies on different MR platforms. This may represent an important barrier, especially for retrospective studies and multicenter projects. Furthermore, while T1WI images show more stability across different experiments, T2WI and especially ADC appear to be more sensitive to changes; thus, different approaches for specific sequences may be needed to increase their stability. Finally, we showed moderate to excellent inter-observer radiomics reproducibility; however, fully automated segmentation pipelines should be implemented to minimize human variability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08282-1.

Funding This study has received funding by NCI U01 CA172320.

Declarations

Guarantor The scientific guarantor of this publication is Bachir Taouli.

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

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained from 16 prospectively recruited patients as part of the NCI U01 CA172320. Written informed consent was waived by the Institutional Review Board for the rest of the cohort.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Data from 16 patients have been previously reported in Bane-2016, Hectors-2016, Hectors-2017, and Jajamovich-2016.

Methodology
- Retrospective
- Observational
- Performed at one institution

References

1. Kumar V, Gu Y, Basu S et al (2012) Radiomics: the process and the challenges. Magn Reson Imaging 30:1234–1248
2. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446
3. Aerts HJWL, Velazquez ER, Leijenaar RTH et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:1–9
4. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762
5. Sullivan DC, Obuchowski NA, Kessler LG et al (2015) Metrology standards for quantitative imaging biomarkers. Radiology 277:813–825
6. Hagiwara A, Fujita S, Ohno Y, Aoki S (2020) Variability and standardization of quantitative imaging: monoparametric to multiparametric quantification, radiomics, and artificial intelligence. Invest Radiol 55:601
7. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B (2020) Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11:1–16
8. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295:328–338
9. Zwanenburg A, Leger S, Vallières M, Löck S (2016) Image biomarker standardisation initiative. arXiv preprint arXiv:161207003
10. Park JE, Kim D, Kim HS et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30:523–536
11. Vallières M, Zwanenburg A, Badic B, Le Rest CC, Visvikis D, Hatt M (2018) Responsible radiomics research for faster clinical translation. Soc Nuclear Med 59:189–193
12. Heus P, Damen JAAG, Pajouheshnia R et al (2018) Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 16:1–12
13. 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
14 Bakr S, Gevaert O, Patel B et al (2020) Interreader variability in semantic annotation of microvascular invasion in hepatocellular carcinoma on contrast-enhanced triphasic CT images. Radiology: Imaging Cancer 2:e190062
15. Echegaray S, Gevaert O, Shah R et al (2015) Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. J Med Imaging 2:041011
16. Baessler B, Weiss K, Pinto Dos Santos D (2019) Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. Invest Radiol 54:221–228
17. Bianchini L, Botta F, Origgi D et al (2020) PETER PHAN: an MRI phantom for the optimisation of radiomic studies of the female pelvis. Physica Med 71:71–81
18. Fiset S, Welch ML, Weiss J et al (2019) Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol 135:107–114
19. Schwier M, van Griethuysen J, Vangel MG et al (2019) Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 9:9441
20. Peerlings J, Woodruff HC, Winfield JM et al (2019) Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep 9:1–10
21. Mahon RN, Hugo GD, Weiss E (2019) Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome. Phys Med Biol 64:145007
22. Bologna M, Corino V, Mainardi L (2019) Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. Med Phys 46:5116–5123
23. Cattell R, Chen S, Huang C (2019) Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art 2:19
24. Yang F, Dogan N, Stoyanova R, Ford JC (2018) Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: a simulation study utilizing ground truth. Phys Med 50:26–36
25. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H (2019) Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. Phys Med Biol 64:165011
26. Ammari S, Pitre-Champagnat S, Dercle L et al (2020) Influence of magnetic field strength on magnetic resonance imaging radiomics features in brain imaging, an in vitro and in vivo study. Front Oncol 10:541663
27. Chernyak V, Fowler KJ, Kamaya A et al (2018) Liver Imaging Reporting and Data System (LI-RADS) version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology 289:816–830
28. O'Sullivan F, Roy S, O'Sullivan J, Vernon C, Eary J (2005) Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET. Biostatistics 6:293–301

29. 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
30. Kessler LG, Barnhart HX, Buckler AJ et al (2015) The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res 24:9–26
31. Hectors SJ, Wagner M, Bane O et al (2017) Quantification of hepatocellular carcinoma heterogeneity with multiparametric magnetic resonance imaging. Sci Rep 7:2452
32. Hectors SJ, Lewis S, Besa C et al (2020) MRI radiomics features predict immuno-oncological characteristics of hepatocellular carcinoma. Eur Radiol 30:3759–3769
33. Chen S, Feng S, Wei J et al (2019) Pretreatment prediction of immunoscore in hepatocellular cancer: a radiomics-based clinical model based on Gd-EOB-DTPA-enhanced MRI imaging. Eur Radiol 29:4177–4187
34. Borhani AA, Catania R, Velichko YS, Hectors S, Taouli B, Lewis S (2021) Radiomics of hepatocellular carcinoma: promising roles in patient selection, prediction, and assessment of treatment response. Abdom Radiol (NY). https://doi.org/10.1007/s00261-021-03085-w
35. Mayerhoefer ME, Szomolanyi P, Jirak D et al (2009) Effects of magnetic resonance image interpolation on the results of texture-based pattern classification: a phantom study. Invest Radiol 44:405–411
36. Bartlett JW, Frost C (2008) Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 31:466–475
37. Saha A, Harowicz MR, Mazurowski MA (2018) Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. Med Phys 45:3076–3085
38. Hu P, Wang J, Zhong H et al (2016) Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget 7:71440
39. van Timmeren JE, Leijenaar RTH, van Elmpt W et al (2016) Test–retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography 2:361