Clinical Imaging 83 (2022) 177–183

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: www.elsevier.com/locate/clinimag

Repeatability, robustness, and reproducibility of texture features on 3 Tesla liver MRI

Vinay Prabhu [a,*], Nicolas Gillingham [a], James S. Babb [a,b], Rahul D. Mali [a,b], Henry Rusinek [a,b], Mary T. Bruno [a], Hersh Chandarana [a,b]

[a] Department of Radiology, NYU Langone Health, New York, NY, United States of America
[b] Center for Advanced Imaging Innovation and Research, NYU Grossman School of Medicine, New York, NY, United States of America

ARTICLE INFO

Keywords:
Texture
Radiomics
Liver MRI

ABSTRACT

Objective: Texture features are proposed for classification and prognostication, with lacking information about variability. We assessed 3 T liver MRI feature variability.
Methods: Five volunteers underwent standard 3 T MRI, and repeated with identical and altered parameters. Two readers placed regions of interest using 3DSlicer. *Repeatability* (between standard and repeat scan), *robustness* (between standard and parameter changed scan), and *reproducibility* (two reader variation) were computed using coefficient of variation (CV).
Results: 67%, 49%, and 61% of features had good-to-excellent (CV $\leq$ 10%) repeatability on ADC, T1, and T2, respectively, least frequently for first order (19–35%). 22%, 19%, and 21% of features had good-to-excellent robustness on ADC, T1, and T2, respectively. 52%, 35%, and 25% of feature measurements had good-to-excellent inter-reader reproducibility on ADC, T1, and T2, respectively, with highest good-to-excellent reproducibility for first order features on ADC/T1.
Conclusion: We demonstrated large variations in texture features on 3 T liver MRI. Further study should evaluate methods to reduce variability.

1. Introduction

Texture features are measures of variation in signal intensity or density within an image. There is increased interest in extending image texture from original 2-dimensional photography to the field of medical image analysis. The main goal is to develop new radiologic texture-based biomarkers for regions of interest (suspicious lesions) on magnetic resonance imaging (MRI) or computed tomography (CT). These metrics fall into groups of varying complexity. The simplest group consists of first order metrics that may be derived from the signal histogram, such as standard deviation or entropy. The more complex measures involve spatial relationship of the signal based on run-lengths and co-occurrence matrices.[1,2]

Texture features have been proposed as predictors of lesion pathology or as prognostic factors. For example, MRI texture features have been proposed as tools to grade bladder cancer,[3] classify liver lesions,[4] and predict biochemical recurrence after radiotherapy for prostate cancer.[5] Texture features also have the potential to be used for prognostication and to detect treatment-related changes. For this purpose, these metrics should be reproducible from reader to reader and robust and repeatable within patients. Furthermore, to take advantage of the big data analytics such as machine learning methods, we need to understand variability in input data which may be acquired not only over time but also when obtained with different parameters and/or at different institutions. However, most published studies are retrospective in nature and utilize imaging performed with different acquisition parameters and signal intensity both within and between studies.[6]

The robustness of texture features has recently been called into question, primarily for CT.[7–9] With respect to MRI texture features, prospective phantom studies have demonstrated sensitivity to acquisition parameters,[10,11] while retrospective studies on in situ normal tissue and lesions have shown sensitivity to acquisition parameters,[12] location,[13] and field strength.[14] To date, there are no prospective studies analyzing the robustness, reproducibility, or repeatability of MRI texture

* Corresponding author at: 660 First Avenue, Third Floor, New York, NY 10016, United States of America.
E-mail addresses: vinay.prabhu@nyulangone.org (V. Prabhu), Nicolas.gillingham@mountsinai.org (N. Gillingham), james.babb@nyulangone.org (J.S. Babb), Rahul.mali@nyulangone.org (R.D. Mali), hr18@nyu.edu (H. Rusinek), mary.bruno@nyulangone.org (M.T. Bruno), hersh.chandarana@nyulangone.org (H. Chandarana).

https://doi.org/10.1016/j.clinimag.2022.01.002
Received 26 October 2021; Received in revised form 9 January 2022; Accepted 12 January 2022
Available online 19 January 2022

features in healthy volunteers. The aim of the present study was to prospectively assess the robustness of commonly employed texture features applied to modern abdominal MRI. We investigated the effect of intra-patient repeatability, parameter change, and inter-reader reproducibility of liver texture features measured repeatedly on healthy volunteers on a single 3 Tesla (3 T) magnet.

2. Materials and methods

This prospective study was HIPAA-compliant and institutional review board-approved.

After obtaining informed consent from five consecutive healthy volunteers (three male, two female, mean age 40.4 years (range 24–66)), abdominal MRI was performed on a 3 T magnet (MAGNETOM Prisma, Siemens, Erlangen, Germany) during May 2018. Sequences were initially obtained using standard clinical institutional protocol parameters for diffusion-weighted imaging (b-values = 0, 800, with apparent diffusion coefficient [ADC] maps) and T2-weighted (T2) and T1-weighted (three-dimensional gradient-echo radial volumetric interpolated breath-hold examination [VIBE], T1) images with spectral fat suppression. Subsequently, to simulate protocol variability in multicenter and retrospective trials, patients were scanned using modified parameters commonly encountered across institutions, vendors, and scanners, and then repeated with baseline parameters ("repeat scan") (Table 1).

Texture analysis was performed using the Radiomics extension in 3DSlicer version 4.8.1 (www.slicer.org).[15] This provides the graphical interface to the open-source code (pyradiomics) for computing texture features.[16] The description of each feature and equations are listed on the pyradiomics website.[17] We included 92 commonly used features, subdivided into classes:

(1) First order (FO, $n = 18$): describe distribution of voxel intensities.
(2) Gray level co-occurrence matrix (GLCM, $n = 23$): describe second-order joint probability for two intensity levels occurring in separate voxels.
(3) Gray level dependence matrix (GLDM, $n = 14$): describe distribution of connected voxels that are within a certain distance and have similar (i.e. within specified range) intensity.
(4) Gray level run length matrix (GLRLM, $n = 16$): quantifies number of consecutive voxels of given intensity.
(5) Gray level size zone matrix (GLSZM, n = 16): consists of connected voxels that share the same intensity.
(6) Neighboring gray tone difference matrix (NGTDM, $n = 5$): quantifies intensity difference between a given voxel and its neighbors.

Two readers (abdominal radiology fellow [VP], fourth year medical student [NG]) with training in 3DSlicer independently placed 5 cm^3 spherical liver regions of interest (ROIs) on axial ADC, T2, and T1 in the right posterior, right anterior, left lateral, and left medial lobes at the level of the main portal vein while avoiding vessels, ducts, or surrounding structures (Fig. 1). This volume was chosen to encompass enough liver tissue and reflect the volume of a typical lesion. Consistent ROI placement across scans was accomplished with software-assisted and visual co-registration. Feature outputs for all four ROIs were averaged for each scan for each patient.

2.1. Statistical analysis

Power calculation was not done for this study to estimate inter- and intra-observer agreement. A proper power analysis requires an a priori specified estimate of the level of agreement. Unfortunately, there is no prior data available to suggest such limits.

Statistics were computed in SAS version 9.4 (SAS Institute, Cary, NC), assessing reliability in terms of:

(1) **Repeatability**: variation between single-reader measurements performed on baseline versus repeat scan using identical parameters
(2) **Robustness**: variation between single-reader measurements for baseline scan and scan acquired after one acquisition parameter was changed
(3) **Reproducibility**: variation between two-reader measurements on the same scan for the same patient

Sources of variation were computed as follows:

(1) Intra-class correlation (ICC): inter-subject variance divided by sum of inter-subject variance and intra-subject variance
(2) Within-subject coefficient of variation (CV): square root of intra-subject variance expressed as a percentage of overall mean

For ICC and CV, restricted maximum likelihood estimation of variance components was used to compute intra- and inter-subject components of the overall variance and these estimates. The CV was only

Table 1
3 T abdominal MRI scan parameters for imaging sequences obtained on each patient. Items bolded with an asterisk (*) are MRI parameters which were changed with respect to standard parameters. VIBE = volume interpolated breath-hold examination.

Imaging sequence	Number of excitations	Time to echo (ms)	Time to repeat (ms)	Flip angle (degrees)	Slice thickness (mm)	Matrix size (pixels^2)
Diffusion-weighted imaging						
Standard	4	85	3000	90	4	192 × 144
Modified 1 (number of excitations)	**2***	85	3000	90	4	192 × 144
Modified 2 (matrix size)	4	85	3000	90	4	**128 × 96***
Modified 3 (slice thickness)	4	85	3000	90	**8***	192 × 144
Repeat scan	4	85	3000	90	4	192 × 144
T2-weighted imaging with fat suppression						
Standard	4	83	5000	120	4	320 × 320
Modified 1 (number of excitations)	**2***	83	5000	120	4	320 × 320
Modified 2 (matrix size)	4	83	5000	120	4	**192 × 192***
Modified 3 (slice thickness)	4	83	5000	120	**8***	320 × 320
Modified 4 (TE)	4	**102***	5000	120	4	320 × 320
Repeat scan	4	83	5000	120	4	320 × 320
T1-weighted imaging with fat suppression (radial VIBE)						
Standard	4	1.58	3.2	12	3	256 × 218
Modified 1 (matrix size)	4	1.58	3.2	12	3	**352 × 300***
Modified 2 (slice thickness)	4	1.58	3.2	12	**5***	256 × 218
Modified 3 (flip angle)	4	1.58	3.2	**6***	3	256 × 218
Repeat scan	4	1.58	3.2	12	3	256 × 218

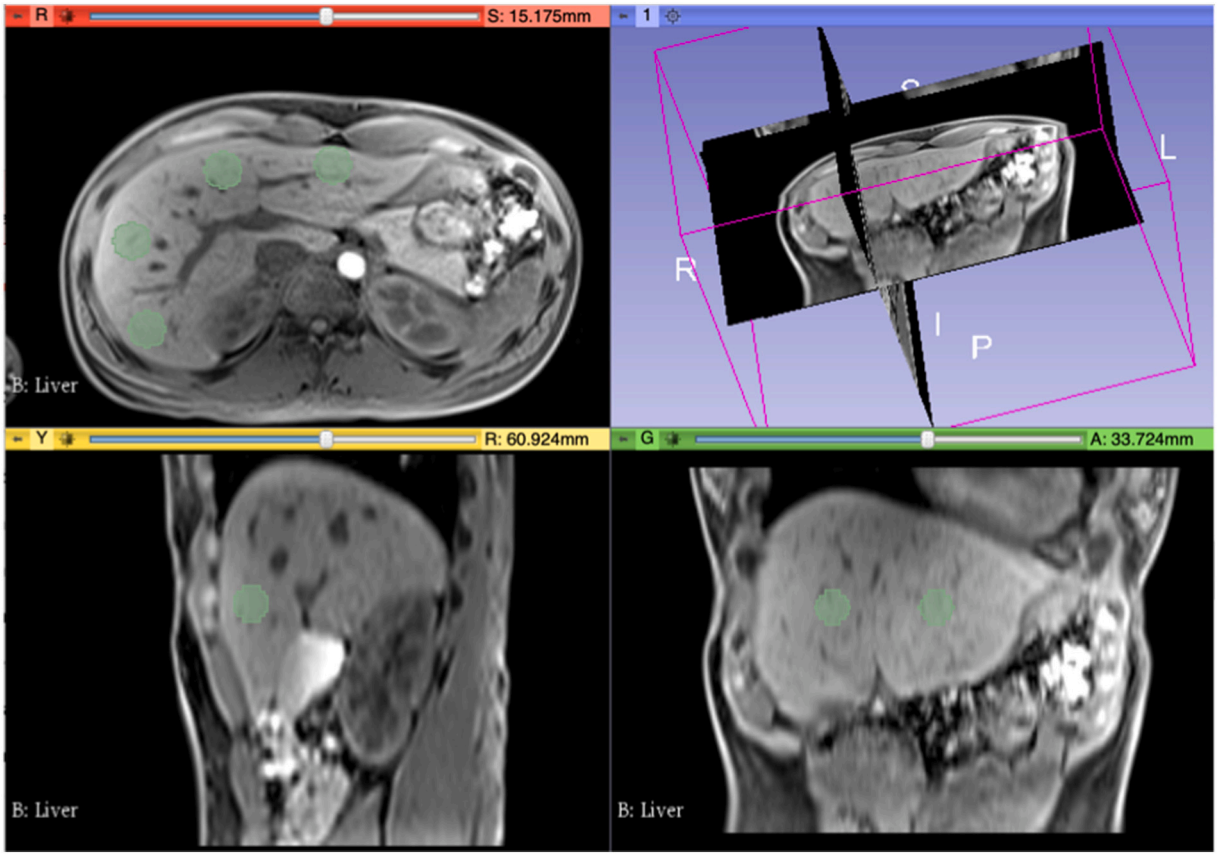

Fig. 1. Example of lesion region of interest (ROI) placement in 3DSlicer on a T1-weighted imaging sequence with fat suppression. 5 cm^3 spherical ROIs were placed in the right lateral, right medial, left lateral, and left medial lobes at the level of the main portal vein.

provided for measures observed to be always positive or always negative. In the latter case, the CV was computed using absolute values of recorded results. Intra-subject differences were computed as the value from reader 1 minus the value from reader 2 (reproducibility) or as the value from the baseline scan minus the value from the second scan (repeatability and robustness). Inter-reader (reproducibility) and inter-scan (repeatability and robustness) reliability were interpreted as poor when ICC < 0.4 or CV $> 20\%$, moderate when $0.4 \leq$ ICC < 0.7 or $10\% <$ CV $\leq 20\%$, good when $0.7 \leq$ ICC < 0.9 or $5\% <$ CV $\leq 10\%$ and excellent when ICC ≥ 0.9 or CV $\leq 5\%$.

To compare scan parameters' impact on texture features, the absolute value of the difference between values derived for a given subject by a given reader from a scan with one parameter changed and the value derived for that subject by that reader from a scan conducted using baseline values of all parameters (i.e., a baseline or repeat scan) was computed for each feature. Since data from five subjects provided minimal power to compare parameters on a per measure basis, this was pooled over measures. Mixed model analysis of variance was used to compare parameters by standardized absolute differences while accounting for lack of statistical independence among differences computed for the same subject. The dependent variable was the vector containing the standardized absolute differences for all texture features from all subjects, both readers, and both baseline scans. An anonymized subject was incorporated as a random classification factor. Pairwise comparisons among the parameters were conducted with the Tukey-Kramer honestly significant difference multiple comparison correction.

3. Results

Five healthy volunteers were scanned. A total of 77 acquisitions were performed (16 per person, except for 13 from one patient who did not complete the three "repeat scans" due to time constraints). Results are summarized below.

3.1. Repeatability

Repeatability was derived by computing texture measures from a 5 cm^3 spherical ROI placed on ADC maps, T1-weighted, and T2-weighted images of the liver. Corresponding texture measures on baseline and repeat scan were compared. Repeatability results are summarized in Supplemental Table 1.

ADC: 59% (54/92) and 67% (60/90) of texture features had good-to-excellent scan-rescan repeatability for both readers by ICC and CV, respectively. GLRLM texture features most frequently had good-to-excellent repeatability for both readers (81% ICC and 63% CV), while FO features least frequently had good-to-excellent repeatability (39% ICC and 35% CV).

T1: 73% (67/92) and 49% (56/89) of texture features had good-to-excellent scan-rescan repeatability for both readers by ICC and CV, respectively. Texture features most frequently with good-to-excellent repeatability for both readers were GLSZM (94%) and GLCM (86%) by ICC and CV, respectively, while FO features least frequently had good-to-excellent repeatability (29% ICC, 19% CV).

T2: 75% (69/92) and 61% (55/90) of texture features had good-to-excellent scan-rescan repeatability for both readers by ICC and CV, respectively. Texture features most frequently with good-to-excellent repeatability for both readers were GLSZM (88%) and GLCM (77%) by ICC and CV, respectively, while FO features least frequently had good-to-excellent repeatability (50% ICC, 29% CV).

3.2. Robustness

The frequency of features reaching threshold CV or ICC levels for each change in MRI acquisition parameter are summarized in Supplemental Table 2 and Figs. 2–3.

For ADC, 41% (113/276) and 22% (59/270) of texture feature measurements had good-to-excellent robustness to parameter changes for both readers by ICC and CV, respectively. For T1, 31% (86/276) and 19% (51/270) of measurements had good-to-excellent robustness to parameter changes for both readers by ICC and CV, respectively. For T2, 42% (156/368) and 21% (77/360) of measurements had good-to-excellent robustness to parameter changes for both readers by ICC and CV, respectively.

The results of our mixed model analysis are shown in Tables 2–3 and Fig. 4. For each texture feature order, the following acquisition parameters were the most robust: slice thickness for FO and flip angle for GLCM, GLDM, GLRLM, GLSZM, and NGTDM (Table 2). The least robust parameter changes were: TE for FO and NGTDM, slice thickness for GLCM, NEx for GLDM and GLRLM, and matrix size for GLSZM. Across all image weightings and texture feature orders, altering the flip angle had the least effect on texture features, while TE had the greatest effect and largest variability as indicated by the widest confidence interval (Fig. 4). Table 3 demonstrates several statistically significant pairwise differences between the effects of parameter changes on texture feature outputs, which are also dependent upon the texture feature group in question.

3.3. Inter-reader reproducibility

Reproducibility results are summarized in Supplemental Table 3, using data from all scans and texture feature measurements.

ADC: 78% (359/460) and 52% (233/450) of individual feature measurements had good-to-excellent inter-reader reproducibility between the two readers according to ICC and CV, respectively. FO texture features had the highest proportion of measurements with good-to-excellent inter-reader reproducibility by ICC (92%) and CV (68%), while GLCM features had the lowest by ICC (67%) and NGTDM the lowest by CV (28%).

T1: 64% (293/460) and 35% (158/450) of individual feature measurements had good-to-excellent inter-reader reproducibility according to ICC and CV, respectively. FO texture features had the highest proportion of measurements with good-to-excellent inter-reader reproducibility by ICC (92%) and CV (61%), while NGTDM features had the lowest by ICC (32%) and CV (4%).

T2: 43% (235/552) and 25% (135/540) of individual texture feature measurements had good-to-excellent inter-reader reproducibility according to ICC and CV, respectively. FO texture features had the highest proportion of measurements with good-to-excellent inter-reader reproducibility by ICC (61%) and GLRLM by CV (33%), while GLCM features had the lowest by ICC (25%) and GLSZM features had the lowest by CV (8%).

4. Discussion

This study prospectively analyzed reproducibility, repeatability, and robustness of texture features in commonly acquired 3 T liver MRI

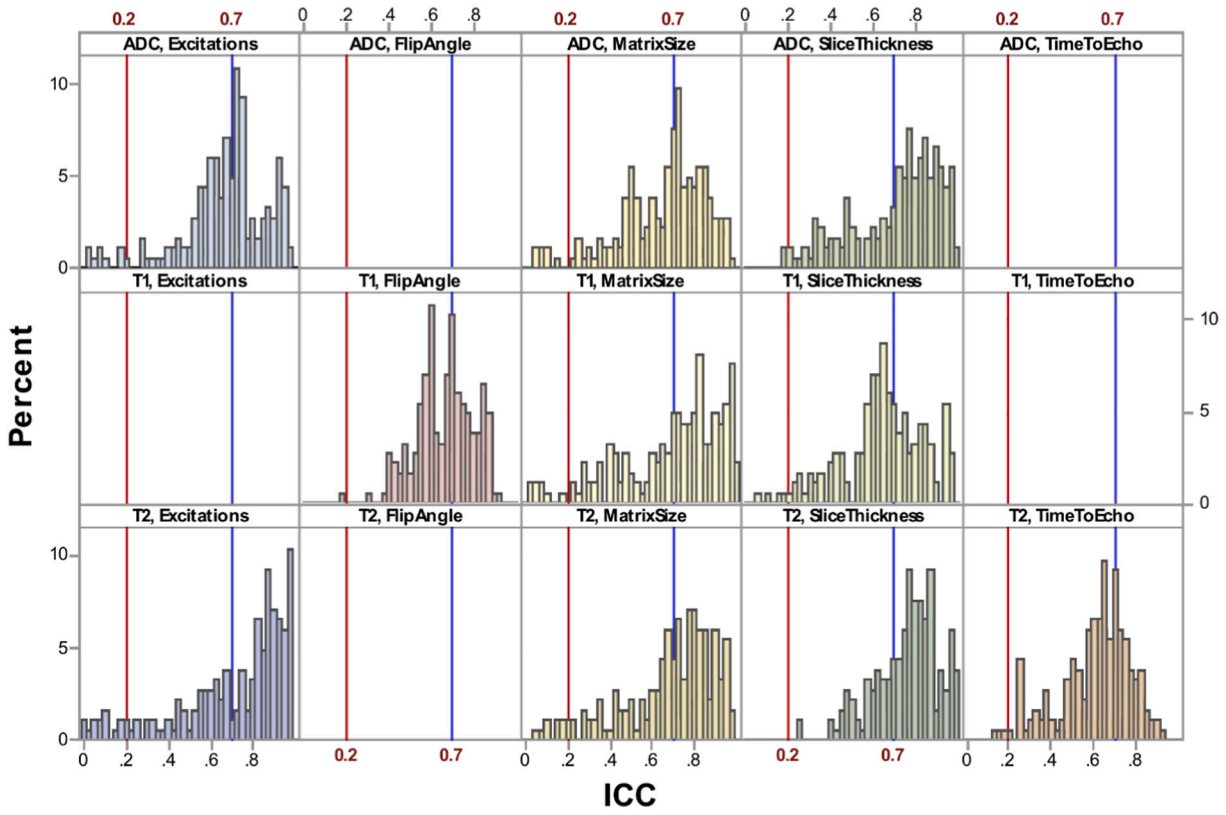

Fig. 2. Histogram of the intra-class correlation (ICC) values as measures of robustness to the change in each scan parameter for the measures associated with each contrast. For each scan histogram ICC values were pooled over measures and readers. Vertical reference lines were added at ICC values of 0.2 (red) and 0.7 (blue) to identify measures with very poor robustness (ICC < 0.2) or good-to-excellent robustness (ICC > 0.7). ADC = apparent diffusion coefficient map, T1 = T1-weighted imaging with fat suppression, T2 = T2-weighted imaging with fat suppression, FO = first order, GLCM = gray level co-occurrence matrix, GLDM = gray level dependence matrix, GLRLM = gray level run length matrix, GLSZM = gray level size zone matrix, NGTDM = neighboring gray tone difference matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

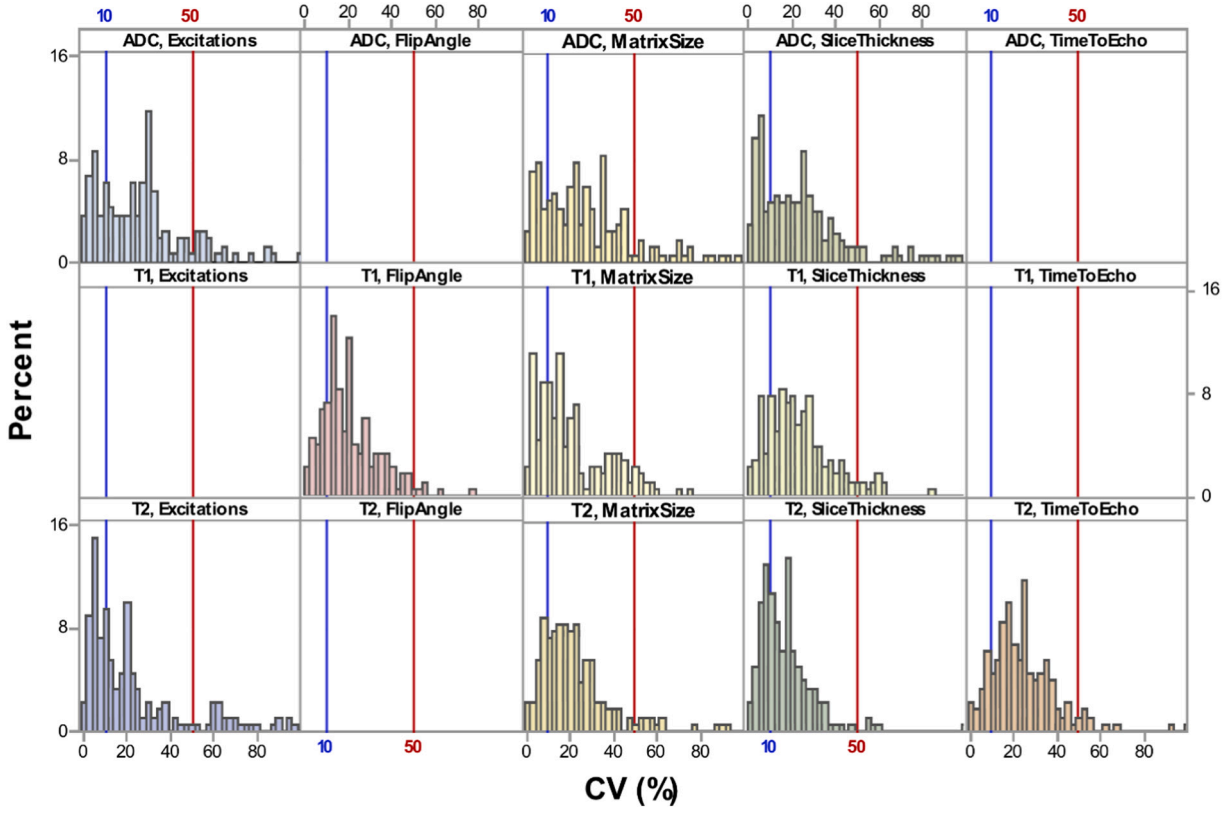

Fig. 3. Histogram of the coefficient of variance (CV) values as measures of robustness to the change in each scan parameter for the measures associated with each contrast. For each scan histogram CV values were pooled over measures and readers. Vertical reference lines were added at CV values of 50% (red) and 10% (blue) to identify measures with very poor robustness (CV > 50%) or good-to-excellent robustness (CV < 10%). ADC = apparent diffusion coefficient map, T1 = T1-weighted imaging with fat suppression, T2 = T2-weighted imaging with fat suppression, FO = first order, GLCM = gray level co-occurrence matrix, GLDM = gray level dependence matrix, GLRLM = gray level run length matrix, GLSZM = gray level size zone matrix, NGTDM = neighboring gray tone difference matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The least squares mean and the standard error (SE) of the least squares mean of the percentage absolute differences associated with each scan parameter among texture features of each order. The least squares mean represents the mean percentage absolute difference adjusted for the effect of texture order and accounting for the lack of statistical independence among differences computed for the same subject. **Bolded** numbers with a ˆ represent the maximum among the means for a given texture feature order and **bolded** numbers with a * represent the minimum among the means for a given texture feature order. That is, the least and most robust parameters for a given texture order are respectively identified by ˆ and *. FO = first order, GLCM = gray level co-occurrence matrix, GLDM = gray level dependence matrix, GLRLM = gray level run length matrix, GLSZM = gray level size zone matrix, NGTDM = neighboring gray tone difference matrix

Scan parameter	FO		GLCM		GLDM		GLRLM		GLSZM		NGTDM	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Excitations	23.5%	2.1%	24.6%	3.5%	**35.6ˆ**	5.5%	**29.5%ˆ**	3.7%	39.4%	5.5%	81.2%	17.6%
Flip angle	22.5%	2.4%	**19.7%***	3.9%	**21.2%***	5.8%	**20.8%***	4.0%	**25.2%***	5.9%	**34.5%***	21.6%
Matrix size	25.9%	2.1%	22.8%	3.3%	32.0%	5.4%	25.5%	3.7%	**40.9%ˆ**	5.4%	55.0%	16.1%
Slice thickness	**18.1%***	2.1%	**27.5%ˆ**	3.3%	28.2%	5.4%	23.1%	3.7%	30.4%	5.4%	47.7%	16.1%
Time to echo	**30.7%ˆ**	2.4%	25.3%	3.9%	27.7%	5.8%	26.1%	4.0%	36.1%	5.9%	**107.2%ˆ**	21.6%

sequences. We found many features to be poorly repeatable among individual patients, have poor robustness to parameter changes, and have limited reproducibility between two readers. Therefore, their use may not be recommended in clinical practice or for longitudinal analysis.

There has been much interest in use of texture features for prognostication and characterization. Recently, studies have analyzed texture features on liver MRI, particularly for identification of hepatic fibrosis[18] and liver lesion characterization.[4] With objective metrics for tissue characterization, it is feasible that texture features could be used in machine learning algorithms for targeted questions.[19,20] For such applications, it is exceedingly important to ensure that results be reliable.

When varying acquisition parameters, we found low concordance rates between feature outputs. Feature measurement *robustness* rates of good-to-excellent were only 31–41% and 19–22%, using ICC and CV, respectively. Our mixed model analysis demonstrated that altering flip angle resulted in the most robust feature outputs across all feature orders aside from FO, and also when grouping all features across all image weightings. On the other hand, robustness to other parameter changes was more variable across different feature orders. Our model demonstrated lower variation in outputs when altering parameters effecting resolution (i.e. matrix size, slice thickness) than those effecting signal-to-noise ratio (i.e. excitations).

Researchers have shown variations in CT texture features when phantoms were scanned using different acquisition parameters and scanners, with intra-CT and inter-CT reproducibility ranging from 42 to

Table 3
P values from the mixed model analysis to compare scan parameters in terms of the mean percentage absolute difference with the Tukey-Kramer honestly significant difference multiple comparison correction. Numbers with a * represent p values less than 0.05. FO = first order, GLCM = gray level co-occurrence matrix, GLDM = gray level dependence matrix, GLRLM = gray level run length matrix, GLSZM = gray level size zone matrix, NGTDM = neighboring gray tone difference matrix.

Scan parameters compared		FO	GLCM	GLDM	GLRLM	GLSZM	NGTDM
Excitations	Flip angle	0.974	0.461	**0.003***	**0.015***	**0.010***	0.246
Excitations	Matrix size	0.423	0.913	0.570	0.207	0.983	0.506
Excitations	Slice thickness	**0.007***	0.694	**0.046***	**0.016***	**0.033***	0.278
Excitations	Time to echo	**0.008***	0.999	0.148	0.620	0.889	0.753
Flip angle	Matrix size	0.295	0.786	**0.018***	0.257	**0.003***	0.850
Flip angle	Slice thickness	0.115	0.077	0.188	0.841	0.578	0.965
Flip angle	Time to echo	**0.008***	0.479	0.426	0.328	0.128	0.068
Matrix size	Slice thickness	**0.001***	0.166	0.419	0.543	**0.005***	0.986
Matrix size	Time to echo	0.073	0.890	0.624	0.999	0.646	0.128
Slice thickness	Time to echo	**0.001***	0.930	1.000	0.650	0.501	0.067

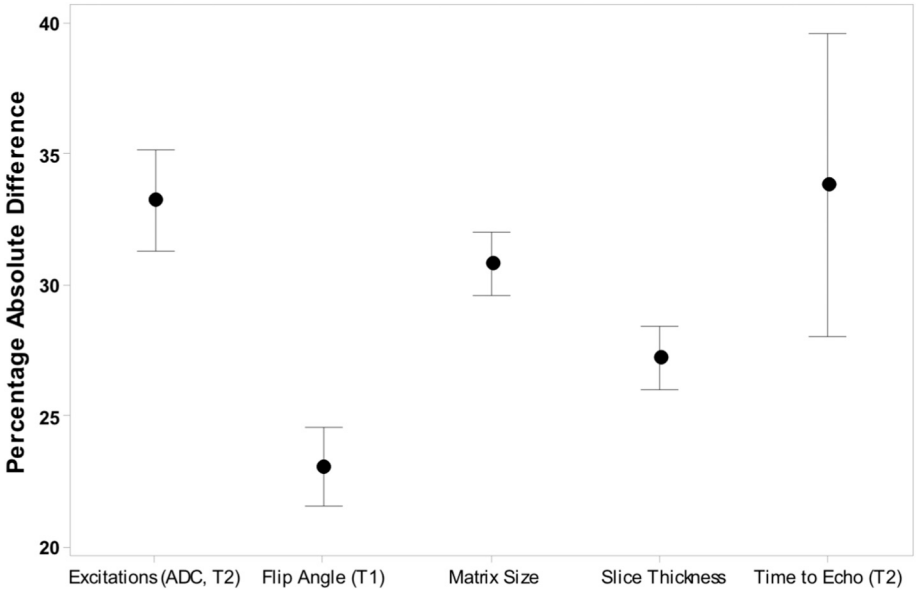

Fig. 4. Interval plot for the percentage absolute differences associated with each scan parameter among texture features of all orders combined. The mean percentage absolute differences are shown as solid circles and the limits of each interval extend from the lower to the upper limit of a 95% confidence interval for the true mean percentage absolute difference. Lower percentage absolute difference (i.e. for flip angle) indicates more robust feature outputs when given parameter changed. ADC = apparent diffusion coefficient map, T1 = T1-weighted imaging with fat suppression, T2 = T2-weighted imaging with fat suppression.

89% and 16–85%, respectively.[8] For MRI, prior study has demonstrated variability in features obtained on brain tumors when varying dynamic range and matrix size, with entropy (an FO feature) the only feature remaining robust,[12] however this study was retrospective and used interpolation post-processing. A study of breast MRI phantoms suggested that resolution may be the most important factor to consider between studies,[11] contrary to our findings. Retrospectively reconstructed brain MR images demonstrated slight differences between features obtained at different slice thickness,[21] as in our study.

We also found differences in features derived from ROIs placed on identical scans by two trained users of a commonly utilized software platform, 3DSlicer. These differences must arise out of slight variation in ROI placement. This result is surprising, as we hypothesized it would be minimal when using large ROIs in normal organs of healthy volunteers. Texture feature *reproducibility* rates of good-to-excellent were only 43–78% and 25–52% using ICC and CV, respectively. That no image-weighting exceeded 78% suggests there is higher than expected inter-reader discordance across all features. This implies an inability of human observers to reproduce x-y-z coordinates of lesion center, and that automated tools may be necessary to ensure reproducibility. Among groups of features, FO features had the highest proportion of good-to-excellent reproducibility across all image weightings. We hypothesize that FO features, being global measures derived from signal distribution over the entire ROI, were the most reproducible between readers because different ROI placement by two readers will cause minimal variability in the shape of the signal histogram on which FO features are based. On the other hand, second order features require more complex and local calculations involving a larger number of variables, which causes greater error propagation,[22] and factors in the differences in signal from neighboring voxels. Because of this, noise-rich MRI images (particularly ADC maps) which accentuate signal heterogeneity will affect second order features more than FO features.

Lastly, we found measurable feature differences on scan-rescan patient images. Similar to reproducibility failures, changes in magnetic field homogeneity induced by different position with respect to coils and patient repositioning could also induce changes in these metrics. Repeatability rates of good-to-excellent were seen in 59%–75% and 49–67% of features using ICC and CV, respectively. In contrast to their high level of reproducibility, FO features were the least repeatable, while GLCM features were the most repeatable across all image weightings. We were surprised to observe that FO features were less repeatable than second order features when ROIs were placed by the same reader on two different scans acquired with identical parameters. While using the same reader on multiple acquisitions may minimize inter-reader error with respect to ROI placement, it may accentuate error from uncontrollable factors such as field variation, patient movement or angulation, and temperature changes. It appears that FO features may be more susceptible to these uncontrollable factors, but additional study is warranted to confirm this finding.

Lack of repeatability, specifically with respect to different patient settings (i.e. different scanners or imaging protocols) has been posited.[13,23] To this end, investigators have developed methods to adjust

for differences on CT, for example using image compensation.[24] Such methodology is less well studied for MRI. In addition, MRI is prone to within- and between-scan intensity changes given that MRI intensity is not standardized to a reference level as Hounsfield units are to water and air. That features were poorly repeatable in patients rescanned on the same day using the same scanner and parameters implies that corrective mechanisms may be difficult to achieve. Limiting studies to a single time point rather than multiple follow-up exams may be a partial mitigator, and using within-patient or within-scan control methods (e.g. normalizing liver lesion data using background liver as a control) could have added benefits. These challenges are compounded for multi-institutional studies where protocols and conditions may differ vastly; for these studies substantial effort should be devoted early on utilizing pilot studies to test texture feature stability and variability prior to selecting tested features and analyzing pooled data.[25] Authors have also suggested harmonization solutions for ensuring reproducibility across scanners and protocol settings in the image domain (e.g. standardized image acquisition, post-processing of raw data, augmentation, style transfer) and feature domain (e.g. normalization and harmonization, including ComBat harmonization).[26]

There are several study limitations. First, we had a small sample size ($n = 5$). However, since each patient was imaged multiple times, the total number of unique imaging sequences for ROI placement was substantially larger ($n = 77$), and is more typical for studies on repeatability and robustness. Second, we analyzed healthy volunteers and studied ROIs placed on normal parenchyma, so our results may not be translatable to lesions, whole organs, or patients with underlying liver disease such as cirrhosis. Since a diseased liver introduces more variability, such as the fraction of ROI containing normal versus fibrotic liver, we should expect an even wider numerical discrepancy of texture features (poorer agreement) in clinical studies. Third, we only studied imaging performed on an MRI from a single vendor. Lastly, texture analysis is sometimes preceded by steps such as signal normalization or discretization; further study should assess measurement robustness on post-processed images.

In conclusion, we demonstrated large variations in texture outputs with respect to intra-patient repeatability, robustness to parameter changes, and inter-reader reproducibility. Further studies are needed to assess reliability of not just individual features but also of multivariate radiomic models that may play an increasing role in coming years. Research should also focus on methods to correct for this variation.

Declaration of competing interest

None.

Acknowledgments

Gautham Sridharan, PhD assisted with data manipulation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinimag.2022.01.002.

References

1. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;SMC-3(6):610–21.
2. Galloway MM. Texture analysis using gray level run lengths. Comput Graphics Image Process 1975;4(2):172–9.
3. Zhang X, Xu X, Tian Q, et al. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. J Magn Reson Imaging Nov 2017; 46(5):1281–8. https://doi.org/10.1002/jmri.25669.
4. Li Z, Mao Y, Huang W, et al. Texture-based classification of different single liver lesion based on SPAIR T2W MRI images. BMC Med Imaging Jul 13 2017;17(1):42. https://doi.org/10.1186/s12880-017-0212-x.
5. Gnep K, Fargeas A, Gutierrez-Carvajal RE, et al. Haralick textural features on T2 -weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. J Magn Reson Imaging Jan 2017;45(1):103–17. https://doi.org/10.1002/jmri.25335.
6. Cattell R, Chen S, Huang C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art Nov 20 2019;2 (1):19. https://doi.org/10.1186/s42492-019-0025-6.
7. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. Mar 24 2016;6:23428. https://doi.org/10.1038/srep23428.
8. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology Aug 2018;288(2):407–15. https://doi.org/10.1148/radiol.2018172361.
9. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology Oct 1 2019:190928. https://doi.org/10.1148/radiol.2019190928.
10. Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study. Med Phys Apr 2009;36(4):1236–43. https://doi.org/10.1118/1.3081408.
11. Waugh SA, Lerski RA, Bidaut L, Thompson AM. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. Med Phys Sep 2011;38(9):5058–66. https://doi.org/10.1118/1.3622605.
12. Molina D, Perez-Beteta J, Martinez-Gonzalez A, et al. Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization. PLoS One 2017;12(6):e0178843. https://doi.org/10.1371/journal.pone.0178843.
13. Chirra P, Leo P, Yim M, et al. Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. J Med Imaging (Bellingham) Apr 2019;6(2):024502. https://doi.org/10.1117/1.JMI.6.2.024502.
14. Whitney HM, Drukker K, Edwards A, Papaioannou J, Giger ML. Robustness of radiomic breast features of benign lesions and luminal A cancers across MR magnet strengths. In: Proceedings of SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis. 105750A; 2018.
15. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging Nov 2012;30 (9):1323–41. https://doi.org/10.1016/j.mri.2012.05.001.
16. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
17. Musunuru HB, Yamamoto T, Klotz L, et al. Active surveillance for intermediate risk prostate cancer: survival outcomes in the sunnybrook experience. J Urol Dec 2016; 196(6):1651–8. https://doi.org/10.1016/j.juro.2016.06.102.
18. House MJ, Bangma SJ, Thomas M, et al. Texture-based classification of liver fibrosis using MRI. J Magn Reson Imaging Feb 2015;41(2):322–8. https://doi.org/10.1002/jmri.24536.
19. Juntu J, Sijbers J, De Backer S, Rajan J, Van Dyck D. Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. J Magn Reson Imaging Mar 2010;31 (3):680–9. https://doi.org/10.1002/jmri.22095.
20. Romeo V, Ricciardi C, Cuocolo R, et al. Machine learning analysis of MRI-derived texture features to predict placenta accreta spectrum in patients with placenta previa. Magn Reson Imaging May 15 2019. https://doi.org/10.1016/j.mri.2019.05.017.
21. Savio SJ, Harrison LC, Luukkaala T, et al. Effect of slice thickness on brain magnetic resonance image texture analysis. Biomed Eng Online Oct 18 2010;9:60. https://doi.org/10.1186/1475-925X-9-60.
22. Farrance I, Frenkel R. Uncertainty of measurement: a review of the rules for calculating uncertainty components through functional relationships. Clin Biochem Rev 2012;33(2):49–75.
23. Hu HT, Shan QY, Chen SL, et al. CT-based radiomics for preoperative prediction of early recurrent hepatocellular carcinoma: technical reproducibility of acquisition and scanners. Radiol Med Aug 2020;125(8):697–705. https://doi.org/10.1007/s11547-020-01174-2.
24. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology Apr 2019;291(1): 53–9. https://doi.org/10.1148/radiol.2019182023.
25. 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. https://doi.org/10.1038/s41598-019-41344-5. 03 18.
26. Mali SA, Ibrahim A, Woodruff HC, et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. J Pers Med Aug 27 2021;11(9). https://doi.org/10.3390/jpm11090842.