FULL PAPER

Repeatability and test–retest reproducibility of mean apparent diffusion coefficient measurements of focal and diffuse disease in relapsed multiple myeloma at 3T whole body diffusion-weighted MRI (WB-DW-MRI)

[1,2]KHALIL ELGENDY, [1,2]TARA D BARWICK, [3,4]HOLGER W AUNER, [3,4]ARISTEIDIS CHAIDOS, [1]KATHRYN WALLITT, [1]ANTONI SERGOT and [1,2]ANDREA ROCKALL

[1]Imperial College Healthcare NHS Trust, London, United Kingdom
[2]The Department of Surgery and Cancer, Imperial College, London, United Kingdom
[3]The Hugh and Josseline Langmuir Centre for Myeloma Research, Centre for Haematology, Department of Immunology and Inflammation, Imperial College London, London, United Kingdom
[4]Department of Haematology, Imperial College Healthcare NHS Trust, London, United Kingdom

Address correspondence to: Mr Khalil ElGendy
E-mail: *khalil.elgendy@nhs.net*

Objective: To assess the test–retest reproducibility and intra/interobserver agreement of apparent diffusion coefficient (ADC) measurements of myeloma lesions using whole body diffusion-weighted MRI (WB-DW-MRI) at 3T MRI.

Methods: Following ethical approval, 11 consenting patients with relapsed multiple myeloma were prospectively recruited and underwent baseline WB-DW-MRI. For a single bed position, axial DWI was repeated after a short interval to permit test–retest measurements.

Mean ADC measurement was performed by two experienced observers. Intra- and interobserver agreement and test–retest reproducibility were assessed, using coefficient of variation (CV) and interclass correlation coefficient (ICC) measures, for diffuse and focal lesions (small ≤10 mm and large >10 mm).

Results: 47 sites of disease were outlined (23 focal, 24 diffuse) in different bed positions (pelvis = 22, thorax = 20, head and neck = 5). For all lesions, there was excellent intraobserver agreement with ICC of 0.99 (0.98–0.99) and COV of 5%. For interobserver agreement, ICC was 0.89 (0.8–0.934) and COV was 17%. There was poor interobserver agreement for diffuse disease (ICC = 0.46) and small lesions (ICC = 0.54).

For test–retest reproducibility, excellent ICC (0.916) and COV (14.5%) values for mean ADC measurements were observed. ICCs of test–retest were similar between focal lesions (0.83) and diffuse infiltration (0.80), while ICCs were higher in pelvic (0.95) compared to thoracic (0.81) region and in small (0.96) compared to large (0.8) lesions.

Conclusion: ADC measurements of focal lesions in multiple myeloma are repeatable and reproducible, while there is more variation in ADC measurements of diffuse disease in patients with multiple myeloma.

Advances in knowledge: Mean ADC measurements are repeatable and reproducible in focal lesions in multiple myeloma, while the ADC measurements of diffuse disease in multiple myeloma are more subject to variation. The evidence supports the future potential role of ADC measurements as predictive quantitative biomarker in multiple myeloma.

INTRODUCTION

Whole body MRI (WB-MRI) has been endorsed by several recent guidelines as an essential imaging modality for patients with multiple myeloma. In the UK, the National Institute for Health and Clinical Excellence (NICE) guideline recommends WB-MRI as one option for first-line imaging for suspected new diagnosis of myeloma.[1] The high sensitivity of WB-MRI has been recognised by the International Myeloma Working Group (IMWG) who recommend it as first-line imaging for asymptomatic myeloma and patients with solitary plasmacytoma.[2] MRI has also been recommended for monitoring treatment response in many subgroups of myeloma patients using qualitative analysis.[3]

Apparent diffusion coefficient (ADC) measurements derived from diffusion-weighted MRI is a potential tool for

objective and functional assessment of disease status and treatment response in many tumours.[4–7] However, translation into clinical practice requires validation of the biomarker through repeatability and reproducibility.[8]

We conducted a literature review to summarise the current evidence of reproducibility of ADC measurements (Table 1). Most of the studies confirmed the reproducibility of ADC measurements, *e.g.* in healthy individuals and in patients with prostate, breast and rectal cancers. In a recently published study, Wennmann et al[10] found good test–retest repeatability of ADC measurements in patients with plasma cell disorders including multiple myelomas. However, there was heterogeneity and inconsistency regarding the methodology used for data acquisition, data analysis and segmentation methodologies in wide variety of pathologies which hinder the building of stronger evidence of the use of ADC measurements.

The current literature suggests a potential role of quantitative ADC measurements in the assessment of treatment response in patients with MM. ADC measurements have been reported to correlate with IMWG criteria for response assessment[18] and could be a potential objective biomarker for response assessment. Myeloma Response Assessment and Diagnosis System (MY-RADS) provide a framework for structured reporting WB-MRI.[19] The Response Assessment Categories (MY-RADS-RACs) are based on objective parameters, (including lesion size, number, and bone marrow signal) and provide a supplementary assessment of treatment response to the standard IMWG response criteria. For the diffuse disease pattern, MY-RADS authors suggest that quantitative ADC measures are not yet practical and therefore not part of the MY-RADS standard. However, a cut-off ADC value of >1400 μm^2/s on post-therapy MRI is used to differentiate between patients likely and highly likely to be responding, but no advice on methods of ADC measurement is provided.

The aim of this study was to assess repeatability and reproducibility of ADC measurements of myeloma lesions on whole body diffusion-weighted MRI (WB-DW-MRI) using 3T MRI.

METHODS AND MATERIALS
Study design
Prospective single centre observational study Institutional review board approval and national research ethics committee approval were obtained (REC reference 14/LO/1833). All patients gave written informed consent.

Patient recruitment and investigations
11 patients with relapsed multiple myeloma requiring systemic therapy were prospectively recruited. Inclusion criteria were age of 18 years or more; confirmed relapsed multiple myeloma (based on IMWG criteria[2]); planned treatment with a licensed novel agent; bone disease visible on conventional imaging (skeletal survey or spinal MRI); and estimated GFR >30 ml/min/ 1.73 m^2. Exclusion criteria included any contraindication to MRI, treatment with any multiple myeloma therapy within the prior 4 weeks, pregnancy and breastfeeding.

All patients underwent baseline WB-DW-MRI. At baseline and for a single bed position, axial DWI was repeated after the patient got off the scanner for a short period of 10 min to permit test and retest DWI measurements. Follow up WB-DW-MRI was performed following two cycles of second-line novel therapy. Treatment response was evaluated based on the International Myeloma Working Group (IMWG) guidelines using serum and urine M protein measurement for six cycles.[19] The novel agents used for second-line therapy included bortezomib, lenalidomide or carfilzomib.

Clinical response assessment
A haematologist, blinded to the research scans, evaluated the clinical responses of the subjects post-cycles 2, 4 and 6 of therapy using IMWG criteria.[20] The response criteria include complete response (CR), very good partial response (VGPR), partial response (PR), minimal response (MR), stable disease (SD) and progressive disease (PD).

MRI acquisition
WB-MRI was performed using Magnetom Verio 3T MRI (Siemens, Erlangen, Germany). All patients were scanned supine with their arms by their sides. Body surface coils were used. DWI sequence parameters included: Transverse orientation, TR: 27600 ms, TE: 65 ms, FoV read: 430 mm, Slice thickness: 5 mm, B-values: 50, 800. Please see Supplementary Table 1 for whole body MRI sequence parameters. Patients were scanned from vertex to upper thighs. ADC maps were generated using a mono-exponential fit using the scanners proprietary software.

Image processing and analysis
The segmentations and measurements of the test and retest were performed on ITK-SNAP (v. 3.6.0) by a single radiologist (KE) with sites of disease checked by a radiology expert (TB) (Figures 1 and 2). The test values were reassessed using Image J (v. 1.5, NIH) to assess the impact of post-processing software on ADC measurements. Second set of segmentation of the test was repeated by the same radiologist (KE) with a 3 week interval between the readings for intraobserver agreement and a second blinded radiologist (AS) for interobserver agreement using ITK-SNAP. The mean ADC, SD and ROI size were recorded using the same software.

Focal lesions were identified as a focal marrow lesion which was hyperintense to background marrow and muscle on b900 s mm^{-2} images, with intermediate ADC and corresponding focal abnormality on DIXON imaging.[19,21] For focal lesions, ROIs were drawn on a single slice with the maximal lesion diameter. In diffuse infiltration, predefined free hand 1.4 cm^2 (47 pixels) ROIs were placed in L5 vertebral body and right and left iliac bones on ADC maps of the pelvis taking care to avoid any focal lesions, bone marrow biopsy tract or artefacts as described previously.[16] In the thoracic bed, 1.4 cm^2 ROI were placed in T3, T4 and T5 vertebra, while in head and neck bed, 1.4 cm^2 were placed in the clivus, arch of C1 and C4. Furthermore, we compared this sampling technique of diffuse infiltration with full segmentation of the ROI in a single slice which was not predefined and was chosen by the reader. For example, we chose the middle part

Table 1. Summary of literature review on test-retest reproducibility of ADC measurements

Author/year of publication	Design, number of participants/ lesions	Assessment criteria	Results summary
Michoux[9] Eur Radiol2021	3T MRI • 24 healthy volunteers	• ROIs in a single slice of the parapspinal muscle, prostate, liver, kidney, spleen, L5 vertebra, posterior iliac crest, femur and white matter (cerebrum) with different sizes.	• CV of ADC was not influenced by the centre or the reader. • Change in ADC must be superior to 66% in lumbar vertebra and 50% in posterior iliac crest and 94% in acetabulum to be significant (other values mentioned for different organs).
Wennmann[10] Invest Radiol 2021	1.5 and 3T MRI • 37 patients with monoclonal plasma cell disorder	• Manual ROIs were placed in BM at posterior iliac crests and muscle tissue using ADC and B800 images. Additional ROI was placed in body of S1 body.	• Maximal CV was 16.2% of the interrater variability and repeatability. • Comparing 1.5 to 3T, larger relative biases of up to −0.526. Normalisation to muscle reduced the bias of T_1W and T_2W but not ADC.
Barrett[11] EJR 2019	3T MRI • 10 prostate cancer patients • Retest same day	• ROIs drawn by consensus of two expert readers on ADC map • ADC histogram analysis MATLAB including media ADC, 10th/90th percentiles, IQR, skewness.	• 10th, 90th centile, and median ADC good repeatability. • Bland–Altman plots showed good repeatability for test and retest analysis for median, percentile and mean range values. • More advanced measures of heterogeneity such as histogram skew, IQR, or mean local range may be limited by their repeatability.
Newitt J[12] Magn Reson Imaging2018	1.5 or 3T MRI • 71 Breast cancer patients • Same day, same imaging session	• Mean and median ADC values were calculated for each composite whole-tumour ROI (using manual segmentation) • 20 cases for intra/inter observer variability.	• ADC repeatability was excellent: wCV = 4.8% (95% CI 4.0, 5.7%), ICC = 0.97 (95% CI 0.95, 0.98), AI = 0.83 (95% CI 0.76, 0.87), and RC = 0.16 * 10−3 mm2/sec (95%CI 0.13, 0.19). • Reproducibility was excellent: interreader ICC = 0.92 (95% CI 0.80, 0.97) and intrareader ICC = 0.91 (95% CI 0.78, 0.96).
Sun[13] Medicine 2017	3T MRI • 26 patients with rectal cancer • 20–30 min between two DWI scans (same session, patient still on the table).	• ADC and IVIM parameters (D, pure diffusion; f, perfusion fraction; D*, pseudodiffusion coefficient) were, respectively, calculated. • ROIs were manually drawn to contour the border of the rectal cancers on the slice (DWI images) with the maximum lesion size • Another circular ROI (100 mm2) was drawn and placed free hand within the left gluteal muscle on the same slice selected above for the first DWI sequence. • The DW-MRI-derived parameters' values were calculated using the pixel-by-pixel fitting method and expressed as the mean values of all the pixels within the ROI	• There was no significant difference in the test and retest values of the DWI-derived parameters (p = .170 [ADC], p = .065 [D], p = .079 [f], and p =.301[D*]). • The test-retest repeatability coefficient for ADC, D, f, and D* was 19.1%, 24.5%, 126.3%, and 197.4%, respectively, greater than the intraobserver values. • ADC and D have better short-term test–retest reproducibility than f and D*. • Considering the poor testintravoxel incoherent motion retest reproducibility for f and D,* variance in these two parameters should be interpreted with caution in longitudinal studies on rectal cancer in which treatment response and recurrence are monitored.
Latifoltojar[14] Eur Radiol 2017	3T MRI • Nine healthy volunteers, 1–11 weeks (median 4 weeks)	• Seven single slice skeletal ROIs (T10 and L4 vertebral bodies, sacroiliac joint and sacral ala, iliac crest, femoral head and neck, mid-shaft femur and distal femur), 2 cm³ circular ROI of the spleen on ADC maps and 3 cm³ circular ROI of subcutaneous adipose tissue at the level of right femoral greater trochanter on mDixon using Osirix	• Bone sFF repeatability was excellent (ICC 0.98) and better than bone ADC (ICC 0.47).
WellerA[15] Eur Radiol 2017	1.5T MRI • 23 patients (30 Malignant lung lesions > 2 cm) • Scanned > 1 h to <1 week	• Whole tumour segmentation using region growing technique (ADEPT) and freehand technique (Osirix) • Assessed lesions that are > 2 cm, and present at least three slices (25 lesions) • whole tumour median ADC (ADCmed) was assessed with Bland–Altman plots	• ADC repeatability coefficient-of-variation is 7.1% for lung tumours > 2 cm. • ADC repeatability coefficient-of-variation is 3.9% for lung tumours > 3 cm. • ADC measurement precision is unaffected by the postprocessing software used. • In multicentre trials, 22% increase in ADC indicates positive treatment response.

(Continued)

Table 1. (Continued)

Author/year of publication	Design, number of participants/lesions	Assessment criteria	Results summary
Messiou C[16] Eur Radiol 2011	• 1.5 T MR • Nine healthy volunteers • FU within 7 days	• 1.3 cm² ROIs were placed in the L5 vertebral body and right and left iliac bones on the ADC map and mean ADC was documented.	• The Bland–Altman limit of reproducibility of mean ADC of bone marrow in normal subjects was 2.0 +/− −86 × 10−6 mm² s⁻¹ • Coefficient of repeatability (r%) expressed as a percentage of the baseline average was 14.8%.
Braithwaite A[17] Radiology 2009	• 3T MRI • 16 healthy volunteers • Mean of 147 days (SD = 2) for follow up scan	• The mean ADCs for three ROIs in five anatomic locations (right hepatic lobe, spleen, and head, body, and tail of pancreas). • The ADC and CV data were then analysed by using repeated-measures analysis of variance	• There were no significant differences in ADCs between imaging sessions 1 and 2. • The mean CV for ADC measurement reproducibility was 14% (95% CI, 13–15%) • Treatment effects of less than approximately 27% (change in ADC divided by pretreatment ADC) will not be clinically detectable with confidence with one acquisition in a single individual

ADC: apparent diffusion coefficient, CV: coefficient of variation, ICC: intraclass correlation coefficient, ROI : region of interest, sFF: signal fat fraction, wCV: within-subject coefficient of variation.

of the vertebral body away from the disc space. For iliac bones, we chose the widest area of the posterior iliac at the level of the sacroiliac joint but discrete from the joints (Figure 3).

On the response assessment studies, the scans were evaluated visually used the MY-RADS-RACS categories[19] by two experienced observers in consensus. In addition, for focal lesions mean ADC measurement of up to five index lesions/patient was documented.[19]

Statistical analysis

Statistical calculations were performed using SPSS (v. 23, (IBM) International Business Machines Corporation). Interclass correlation coefficients (ICC) estimates along with their 95% confidence intervals (CIs) were calculated using a two-way random, absolute agreement, single measure model with 95% CI for the mean ADC values for all, focal lesions and diffuse infiltration. ICC values less than 0.5 suggest poor agreement, 0.5–0.75 moderate agreement, 0.75–0.9 good and greater than 0.9 excellent agreement.[22] In addition, coefficients of variation (%) were calculated using MedCalc Statistical Software (v. 14.8, Belgium). The same software was used to generate Scatter plots (with line of equality) and Bland–Altman plots (difference *vs* means).

RESULTS

11 patients with relapsed multiple myeloma were recruited. Patient demographics are summarised in Table 2. A total of 47 regions of disease were identified (23 focal, 24 diffuse).

Table 3 summarises the values of ICC and CV and Figure 4 and Supplementary Figures 1–3 show the scatter and Bland–Altman plots for of test–retest reproducibility and intra- and interobserver agreements. Comparisons were made between ADC measurements for diffuse disease and focal lesions, lesion location (thoracic and pelvic bed positions) and lesion size (small and large lesions). Comparison is also made between different techniques for assessment of diffuse infiltration, *i.e.* sampling *vs* segmentation techniques.

Overall, there was excellent intraobserver agreement with ICC being 0.99 and CV being 5% (*n* = 47) (Figure 4). The interobserver agreement was good with lower value of ICC (0.89) and higher value of CV (17%). The test–retest reproducibility had excellent ICC and CV values (0.916 and 14.5% respectively). Similar values of ADC measurements were obtained between the two software: Image J (v. 1.5, NIH) and ITK SNAP (v. 3.6.0) and no further statistical analysis were required.

Diffuse disease *vs* focal lesions

Focal lesion ADC measurements (*n* = 23) had excellent intra- and interobserver agreements and test–retest reproducibility with ICCs values above 0.8 and CV values below 15% (Supplementary Figure 1). However, the ADC measurements for diffuse infiltration (*n* = 24) using sampling technique had a poor interobserver agreement (ICC = 0.46, CV = 29%) and moderate test–retest reproducibility (ICC = 0.81, and CV = 19.1%). On repeating ADC measurements of diffuse disease using whole slice segmentation technique, there was improvement in interobserver agreement

Figure 1. Test (a) and Retest (b) images of a focal lesion in the right posterior iliac bone (b900 and ADC, with segmentation using ITK SNAP software). ADC, apparent diffusion coefficient.
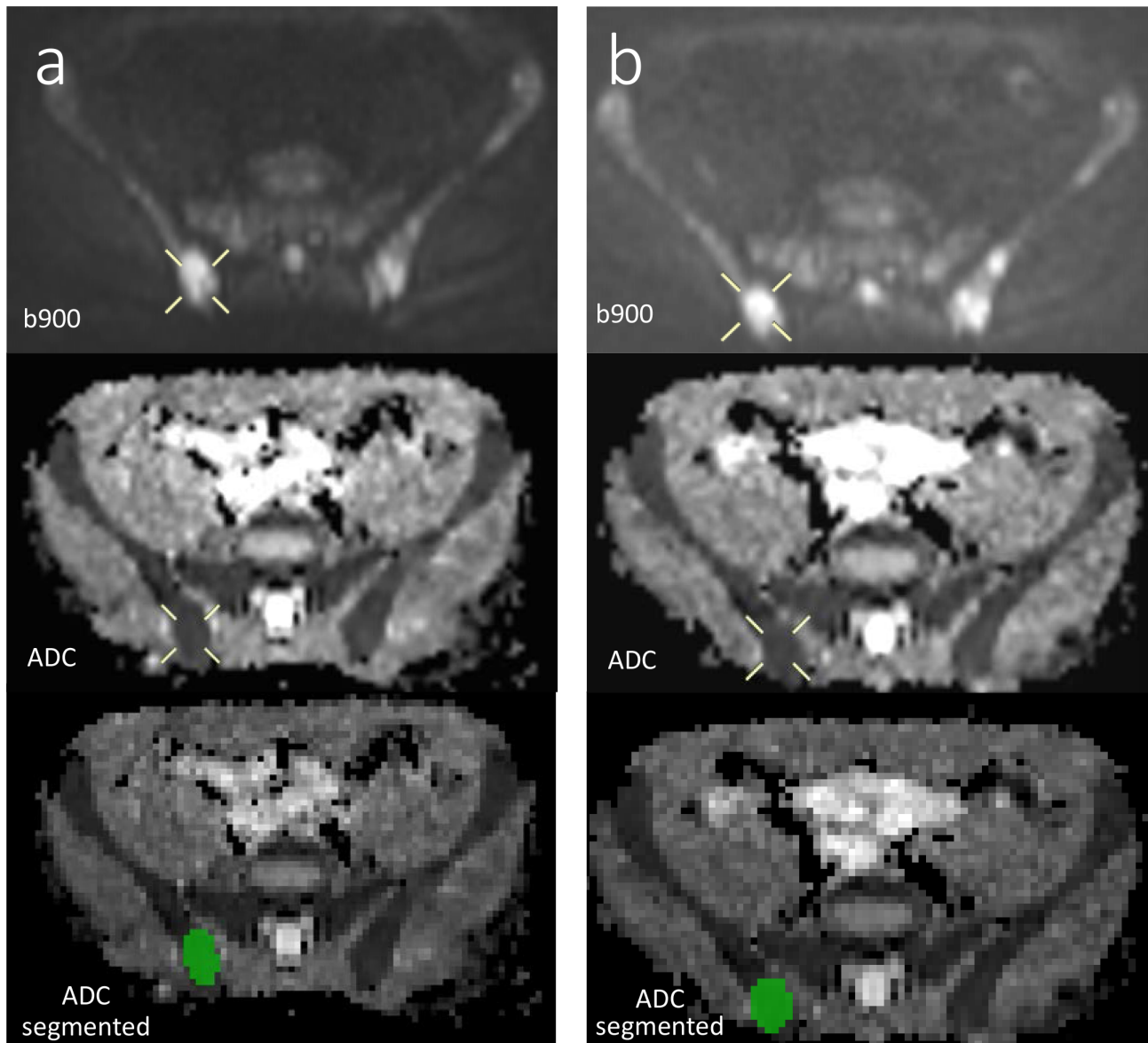
(ICC = 0.9 *vs* 0.46) in contrast to test–retest reproducibility that was moderate (ICC = 0.58 *vs* 0.81; CV = 27.5%).

Impact of bed position

For focal lesions, the intraobserver agreement was excellent for both pelvic and thoracic bed positions (Supplementary Figure 2). The interobserver agreements were good for pelvic focal lesions and moderate for thoracic focal lesions. Higher ICC values were achieved for focal pelvic compared to focal thoracic lesions in test–retest reproducibility (ICC = 0.94, CV 9.2% for focal pelvic lesions *vs* 0.6, CV 13.8% for focal thoracic lesions). Head and neck lesions (n = 5) showed excellent intra- and interobserver agreement and test–retest reproducibility. However, the number of lesions are too small to make any meaningful conclusions or comparisons.

Small *vs* large lesions

Excellent intraobserver agreement was achieved for both small (<10 mm, n = 8) and large (>10 mm, n = 15) focal lesions (Supplementary Figure 3). The test–retest reproducibility was excellent for small lesions and moderate for large lesions. For small lesions, there was moderate interobserver agreement with ICC value of 0.54 and CV of 18%. For large lesions, interobserver agreement was excellent (ICC = 0.9 CV=9.8%).

DISCUSSION

WB-DW- MRI is now considered standard of care for imaging of multiple myeloma patients and increasingly used for response assessment. The recent Myeloma Response Assessment and Diagnosis System[19] guidelines propose visual response assessment categories. However, they also stipulate ADC cut-offs to

Figure 2. Test (a) and Retest (b) images of focal expansile lesion in the right clavicle (b900 and ADC, with segmentation using ITK SNAP software). ADC, apparent diffusion coefficient.
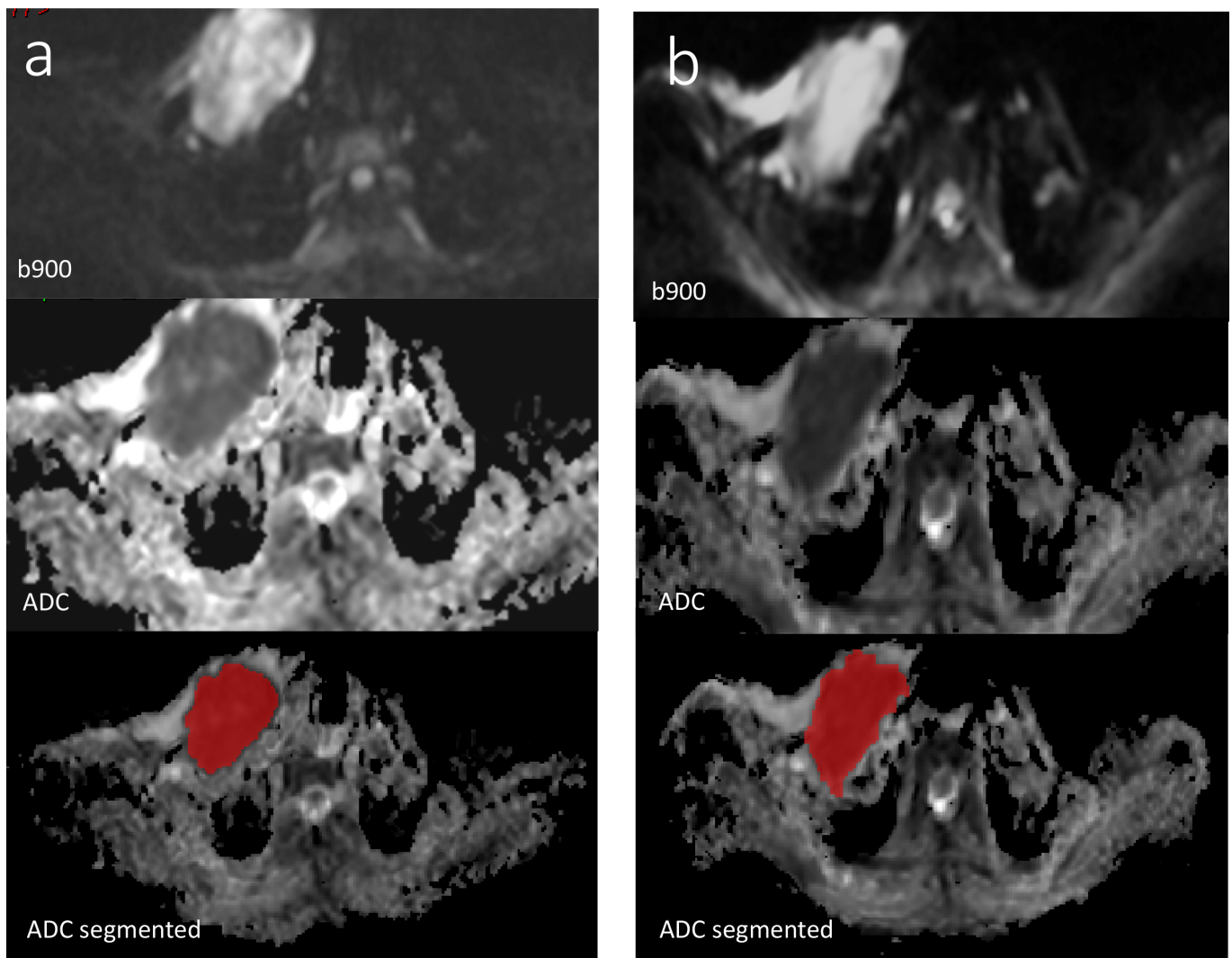
allocate into various response categories.[19] In addition, there is increasing interest in assessing change in ADC measurement as a biomarker of response. Therefore, knowledge of repeatability and reproducibility of ADC measurement is highly important.

In this study, there was excellent test–retest reproducibility (ICC = 0.916, CV = 14.5%) and repeatability in the form of intraobserver agreement (ICC = 0.99, CV = 5%) and to a lesser degree interobserver agreement (ICC = 0.89, CV = 17.9%) for all lesions. When considering focal lesions, intraobserver agreement is excellent with moderate interobserver agreement and test–retest reproducibility. The ADC measurement of focal lesions in the thoracic bed position was more subject to variation than the pelvic bed position. This may be due to the thoracic bed position being more subject to movement, in addition to a greater potential for different slices being selected on the retest imaging.

As demonstrated from the summary of literature in Table 1, ADC measurements have been reported to be repeatable and reproducible in healthy tissues[17] and prostate,[11] breast,[12] lung[15]

and rectal cancers[13] with similar values achieved in our study in multiple myeloma lesions. In a recent prospective study, Wennmann et al.,[10] assessed repeatability and reproducibility of ADC measurements of pelvic bone marrow in patients with monocloncal plasma cell disorders. Overall, CoV for pelvic ADC measurement was 14.5% for test–retest reproducibility at 1.5T in 27 subjects and 15.8% for interobserver agreement using combined data at 1.5 and 3T. Similar values were achieved in our study for all lesions (14.5 and 17.9% respectively) and also in pelvic lesions (15 and 22.8% respectively). Reproducibility of ADC measurements between 1.5 and 3T showed very high CV of 41.3% for pelvic ADC measurement which they postulated was due to susceptibility differences between trabecular calcium hydroxyapatite and bone marrow being more marked at higher magnetic field strengths. Unlike our study, test–test reproducibility at 3.0 T was not assessed in their study and in addition focal lesions were not assessed. Assessment of focal lesions is important as the emphasis of ADC measurement in MY-RADS response is with focal lesions. These findings separately confirm

Figure 3. Comparison between two methods of segmentation of diffuse disease within the pelvis and lumbar spine using fixed sampling technique (a, b) and full segmentation technique (c, d) in a single slice at posterior iliac (a, c) and L5 vertebral body (b, d). ADC, apparent diffusion coefficient.
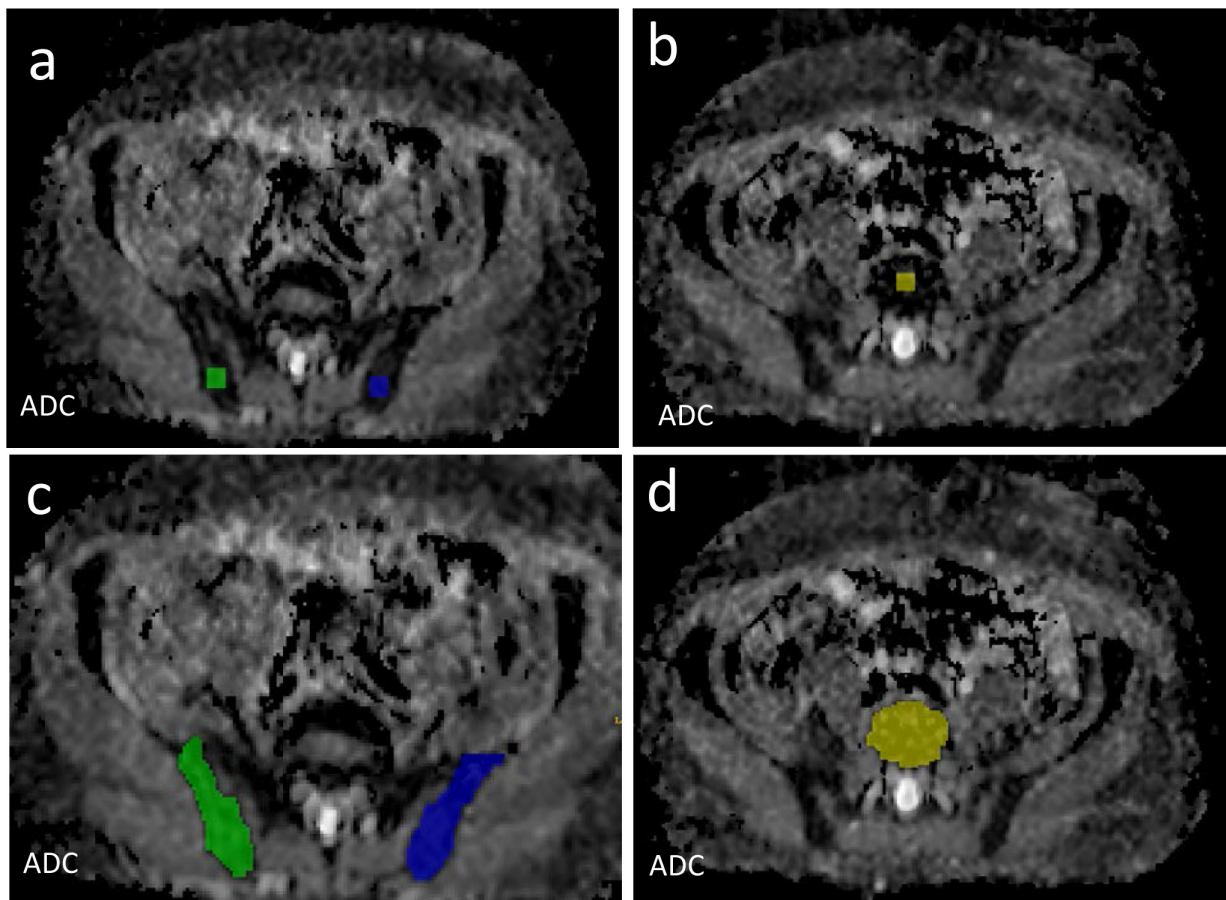

Table 2. Patient demographics

Gender	
Male	10
Female	1
Mean age (years, range)	59.5 (45–71)
Imaging patterns	
Focal	3
Diffuse	5
Focal on diffuse	3
Myeloma subtype	
IgG	8
IgA	3
Novel agent treatment	
Lenalidomide-based	7
Bortezomib-based	3
Carfilzomib-based	1

the test–retest reproducibility of the ADC measurements in patients with multiple myeloma.

The size of lesions has been reported to impact the repeatability of ADC measurement in a number of tumours including myeloma.[21,23,24] Weller et al (2011) demonstrated that ADC measurement variability is lower for large size lung lesions in comparison to small lung lesions (>3 cm, CV 3.9%;<3 cm, CV = 9.6%).[15] Only one previous retrospective study has assessed the impact of lesion size in myeloma ADC measurements at 1.5T, Barwick et al[2121] reported for mean ADC, excellent ICC and low CV for inter- and intraobserver agreement for small (<10 mm) and large (>10 mm) lesions. They did not assess test–retest reproducibility. In our study, for large lesions interobserver agreement was excellent but reduced for smaller lesions (ICC 0.9, CV 9.8% large *vs* ICC 0.54, CV 18% small) which we postulate may be due to smaller lesions being more subject to partial voluming and noise making them more difficult to outline.

The choice of ROI has a major impact on the ADCs values and its repeatability and reproducibility.[23] Blazic et al[5] in a rectal cancer study concluded that the larger measurement methods yield greater accuracy in response assessment. However,

Table 3. Summary of ICC and CV for intraobserver agreement, interobserver agreement and test–retest reproducibility

		Intraobserver agreement			Interobserver agreement			Test–retest		
		ICC		CV	ICC		CV	ICC		CV
Overall	n = 47	0.99	(0.98–0.99)	5%	0.89	(0.8–0.934)	17.90%	0.916	(0.85–0.95)	14.50%
Diffuse (S1*)	n = 24	0.98	(0.95–0.99)	5.80%	0.46	(0.06–0.72)	29%	0.81	(0.59–0.9)	19.10%
Diffuse (S2*)	n = 24	0.95	(0.85–0.98)	9.6%	0.9	(0.55–0.96)	14%	0.58	(0.23–0.8)	27.5%
Focal	n = 23	0.98	(0.95–0.99)	4.50%	0.82	(0.63–0.92)	12.80%	0.83	(0.65–0.925)	12.10%
Impact of bed position										
Thorax	n = 20	0.97	(0.93–0.99)	5.69%	0.87	(0.7–0.95)	13.90%	0.81	(0.59–0.92)	14.10%
T-Diffuse	n = 6	0.99	(0.97–0.99)	2%	0.54	(-0.06–0.9)	31%	0.73	(0.03–0.96)	13.9%
T-Focal	n = 14	0.91	(0.74–0.97)	6%	0.71	(0.32–0.89)	11.6%	0.6	(0.11–0.85)	13.8%
Pelvis	n = 22	0.99	(0.99–0.99)	3.48%	0.89	(0.75–0.95)	22.80%	0.953	(0.89–0.98)	14.96%
P-Diffuse	n = 15	0.97	(0.93–0.99)	6.40%	0.33	(-0.18–0.71)	32.70%	0.77	(0.44–0.92)	22.90%
P-Focal	n = 7	0.99	(0.99–0.99)	0.90%	0.86	(0.42–0.97)	14.80%	0.94	0.83–0.99	9.20%
Head and Neck	n = 5	0.988	(0.91–0.99)	5.5%	0.93	(0.5–0.99)	11.4%	0.93	(0.47–0.99)	12.9%
H-Diffuse	n = 3	0.95	(0.19–0.99)	7.5%	0.86	(-1.2–0.99)	12.9%	0.77	(-0.56–0.99)	16.8%
H-Focal	n = 2	0.99	(0.97–0.99)	ID**	0.98	(0.13–0.99)	ID**	0.99	(0.36–0.99)	ID**
Impact of lesion size										
Small	n = 8	0.98	(0.91–0.99)	3.10%	0.54	(-0.15–0.89)	18%	0.96	(0.81–0.99)	4.50%
Large	n = 15	0.98	(0.93–0.99)	4.92%	0.9	(0.73–0.96)	9.82%	0.8	(0.51–0.93)	13.50%

CV, coefficient of variation; ICC, intraclass correlation coefficient.
* Diffuse S1: Sampling technique, S2: Whole Segmentation technique. ID** : Insufficient

Figure 4. Bland–Altman plots (a,c,e) and scatter plotgrams (b, d, f) for mean ADC values of overall test– retest (a, b), interobserver agreement (c, d) and intraobserver agreement (e, f). ADC, apparent diffusion coefficient.
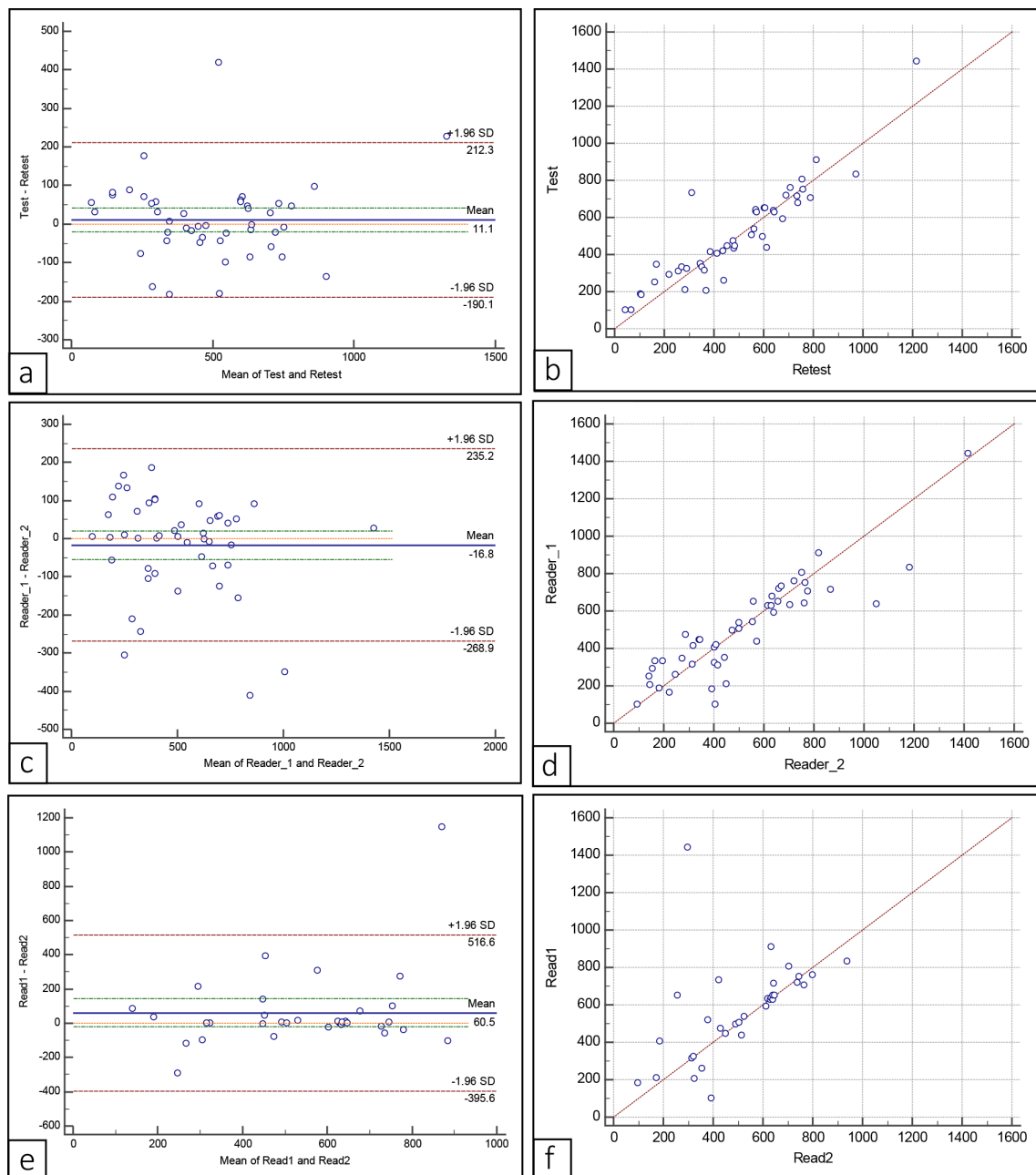
Nogueira et al[25] showed that smaller fixed ROIs have higher ADC reproducibility and less variability than segmenting the whole lesion in primary breast tumours. For focal lesions, we adopted single slice ROI at maximum axial dimension of the tumour as opposed to whole tumour multislice outlining which is a potential limitation. However, this is the approach taken in previous myeloma studies[14] since unlike other primary tumours, myeloma lesions tend to be numerous so a single slice approach is less time consuming which may be more feasible for potential future clinical use. However, future studies may assess whole tumour segmentation tools using machine learning algorithm to measure disease burden and assess response which the current evidence supports its feasibility[26,27] This is currently under

ongoing research in our institute (Machine Learning in Myeloma Response (MALIMAR) study.[28]

For diffuse disease, we used two different methods: fixed ROI sampling techniques and segmentation of whole area of interest on a single slice. Both had excellent intraobserver agreement. However, the first technique of predefined ROI had better test–retest reproducibility (ICC = 0.81) but poorer interobserver agreement (ICC = 0.46) in comparison to the second whole segmentation technique (ICC = 0.58 and 0.9 respectively) (Table 3). It is interesting that with whole slice segmentation test–retest reproducibility was inferior to single slice fixed size ROI which may reflect reduced precision of ADC measurement

at the boundaries of bone marrow with other tissue when whole segments are outlined manually. Whole slice segmentation is time-consuming in clinical practice so automated approaches are desirable. Recently published work by Wennmann et al, demonstrated that a deep learning algorithm can perform automated bone marrow segmentation of 30 different bones from which automated extraction of ADC values for whole bones can be performed. This method could lead to improvements in reproducibility of ADC measurements.[10]

Messiou et al, assessed test–retest reproducibility of ADC measurement of bone marrow in non-diseased healthy volunteers at 1.5 T[16] and reported better values to the assessment of diffuse disease in the pelvis in comparison to the results in our study (CV = 14.8% *vs* 32%). The heterogeneity of ADC measurements in diffuse disease can be explained by the increased likelihood of selecting a different slice/ region for 'diffuse' as opposed to 'focal' lesions. In addition, the marrow of healthy volunteers may potentially be less heterogeneous than diseased marrow. Further focused research with larger power is required before drawing any specific conclusion about the best method for assessment of diffuse disease.

DWI and ADC has significantly improved correlation of the imaging with clinical and laboratory measurements[29] with accurate reflection of the disease course and treatment responses.[30] However, the evidence is scarce with regards ADC prediction of clinical response to treatment. In a recent prospective study, Michoux et al[9] suggested that clinically significant changes in ADC must be greater than 50% (posterior iliac crest), 66% (L5 vertebra), 68% (femur) and 94% (acetabulum). Wu et al[31] illustrated that by using ADC value of 1×10^{-3} mm^2/s, ADC has positive predictive value of deep response of 60%. Zhang et al used 0.81×10^{-3} mm^2/s as cut-off and found that ADC has sensitivity of 54% and specificity of 68% of predicting increased ADC in response to treatment.[6] These results support our future research efforts in understanding the potential role of ADC measurements as predictive quantitative biomarker in multiple myeloma patients.

There are limitations to our study. It is a single centre using a single MRI machine. Further studies are needed across different MRI scanners from different vendors. Also, validation studies are needed to reach conclusions regarding the ADC values which is affected by different DWI protocols and MRI scanners. The study recruited 11 patients over a period of 2 years which is a relatively small number. However, there were 47 focal lesions and diffuse infiltration that was studied which allowed subgroup analysis of the effect of the bed site and size on the variability. The prospective design of the study and blinding of readers to each other and to the repeat data are also among the strengths of the current project.

In conclusion, mean ADC measurements at 3T are repeatable and reproducible in focal lesions in multiple myeloma patients. The measurement of diffuse disease is more subject to variation. The evidence supports future research of the role of ADC measurements as a potential objective tool in assessment of disease status, response to interventions and prognosis in multiple myeloma patients.

REFERENCES

1. NICE. *Myeloma: diagnosis and management*. NICE Guidelines; 2016.

2. Dimopoulos M, Terpos E, Comenzo RL, Tosi P, Beksac M, Sezer O, et al. International myeloma working group consensus statement and guidelines regarding the current role of imaging techniques in the diagnosis and monitoring of multiple myeloma. *Leukemia* 2009; **23**: 1545–56. https://doi.org/10.1038/leu.2009.89

3. Chantry A, Kazmi M, Barrington S, Goh V, Mulholland N, Streetly M, et al. Guidelines for the use of imaging in the management of patients with myeloma. *Br J Haematol* 2017; **178**: 380–93. https://doi.org/10.1111/bjh.14827

4. Jajodia A, Mahawar V, Chaturvedi AK, Rao A, Singla R, Mitra S, et al. Role of ADC values in assessing clinical response and identifying residual disease post-chemo radiation in uterine cervix cancer. *Indian J Radiol Imaging* 2019; **29**: 404–11. https://doi.org/10.4103/ijri.IJRI_339_19

5. Blazic IM, Lilic GB, Gajic MM. Quantitative assessment of rectal cancer response to neoadjuvant combined chemotherapy and radiation therapy: comparison of three methods of positioning region of interest for ADC measurements at diffusion-weighted MR imaging. *Radiology* 2017; **282**(2): 615. https://doi.org/10.1148/radiol.2017164040

6. Zhang Y, Xiong X, Fu Z, Dai H, Yao F, Liu D, et al. Whole-body diffusion-weighted MRI for evaluation of response in multiple myeloma patients following bortezomib-based therapy: A large single-center cohort study. *Eur J Radiol* 2019; **120**: 108695. https://doi.org/10.1016/j.ejrad.2019.108695

7. Philippe J, Jochen F, Mathias S, Günther S, Christian R, Arno B, et al. Diffusion-weighted MRI improves response assessment after definitive radiotherapy in patients with NSCLC. *Cancer Imaging* 2021.

8. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 2017; **14**: 169–86. https://doi.org/10.1038/nrclinonc.2016.162

9. Michoux NF, Ceranka JW, Vandemeulebroucke J, Peeters F, Lu P, Absil J, et al. Repeatability and reproducibility of ADC measurements: a prospective multicenter whole-body-MRI study. *Eur Radiol* 2021; **31**: 4514–27. https://doi.org/10.1007/s00330-020-07522-0

10. Wennmann M, Thierjung H, Bauer F, Weru V, Hielscher T, Grözinger M, et al. Repeatability and reproducibility of ADC measurements and MRI signal intensity measurements of bone marrow in monoclonal plasma cell disorders: A prospective bi-institutional multiscanner, multiprotocol study. *Invest Radiol* 2022; **57**: 272–81. https://doi.org/10.1097/RLI.0000000000000838

11. Barrett T, Lawrence EM, Priest AN, Warren AY, Gnanapragasam VJ, Gallagher FA, et al. Repeatability of diffusion-weighted MRI of the prostate using whole lesion ADC values, skew and histogram analysis. *Eur J Radiol* 2019; **110**: 22–29. https://doi.org/10.1016/j.ejrad.2018.11.014

12. Newitt DC, Zhang Z, Gibbs JE, Partridge SC, Chenevert TL, Rosen MA, et al. Test-retest repeatability and reproducibility of ADC measures by breast DWI: results from the ACRIN 6698 trial. *J Magn Reson Imaging* 2019; **49**: 1617–28. https://doi.org/10.1002/jmri.26539

13. Sun H, Xu Y, Xu Q, Shi K, Wang W. Rectal cancer. *Medicine* 2017; **96**: e6866. https://doi.org/10.1097/MD.0000000000006866

14. Latifoltojar A, Hall-Craggs M, Bainbridge A, Rabin N, Popat R, Rismani A, et al. Whole-body MRI quantitative biomarkers are associated significantly with treatment response in patients with newly diagnosed symptomatic multiple myeloma following bortezomib induction. *Eur Radiol* 2017; **27**: 5325–36. https://doi.org/10.1007/s00330-017-4907-8

15. Weller A, Papoutsaki MV, Waterton JC, Chiti A, Stroobants S, Kuijer J, et al. Diffusion-weighted (DW) MRI in lung cancers: ADC test-retest repeatability. *Eur Radiol* 2017; **27**: 4552–62. https://doi.org/10.1007/s00330-017-4828-6

16. Messiou C, Collins DJ, Morgan VA, Desouza NM. Optimising diffusion weighted MRI for imaging metastatic and myeloma bone disease and assessing reproducibility. *Eur Radiol* 2011; **21**: 1713–18. https://doi.org/10.1007/s00330-011-2116-4

17. Braithwaite AC, Dale BM, Boll DT, Merkle EM. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. *Radiology* 2009; **250**: 459–65. https://doi.org/10.1148/radiol.2502080849

18. Paternain A, García-Velloso MJ, Rosales JJ, Ezponda A, Soriano I, Elorz M, et al. The utility of ADC value in diffusion-weighted whole-body MRI in the follow-up of patients with multiple myeloma. correlation study with [18]f-FDG PET-CT. *Eur J Radiol* 2020; **133**: 109403. https://doi.org/10.1016/j.ejrad.2020.109403

19. Messiou C, Hillengass J, Delorme S, Lecouvet FE, Moulopoulos LA, Collins DJ, et al. Guidelines for acquisition, interpretation, and reporting of whole-body MRI in myeloma: myeloma response assessment and diagnosis system (MY-RADS). *Radiology* 2019; **291**: 5–13. https://doi.org/10.1148/radiol.2019181949

20. Kumar S, Paiva B, Anderson KC, Durie B, Landgren O, Moreau P, et al. International myeloma working group consensus criteria for response and minimal residual disease assessment in multiple myeloma. *Lancet Oncol* 2016; **17**: e328–46. https://doi.org/10.1016/S1470-2045(16)30206-6

21. Barwick T, Orton M, Koh DM, Kaiser M, Rockall A, Tunariu N, et al. Repeatability and reproducibility of apparent diffusion coefficient and fat fraction measurement of focal myeloma lesions on whole body magnetic resonance imaging. *Br J Radiol* 2021; **94**(1120): 20200682. https://doi.org/10.1259/bjr.20200682

22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; **15**: 155–63. https://doi.org/10.1016/j.jcm.2016.02.012

23. Lambregts DMJ, Beets GL, Maas M, Curvo-Semedo L, Kessels AGH, Thywissen T, et al. Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability. *Eur Radiol* 2011; **21**: 2567–74. https://doi.org/10.1007/s00330-011-2220-5

24. Ma C, Guo X, Liu L, Zhan Q, Li J, Zhu C, et al. Effect of region of interest size on ADC measurements in pancreatic adenocarcinoma. *Cancer Imaging* 2017; **17**: 13. https://doi.org/10.1186/s40644-017-0116-6

25. Nogueira L, Brandão S, Matos E, Nunes RG, Ferreira HA, Loureiro J, et al. Region of interest demarcation for quantification of the apparent diffusion coefficient in breast lesions and its interobserver variability. *Diagn Interv Radiol* 2015; **21**: 123–27. https://doi.org/10.5152/dir.2014.14217

26. von Brandis E, Jenssen HB, Avenarius DFM, Bjørnerud A, Flatø B, Tomterstad AH, et al. Automated segmentation of magnetic resonance bone marrow signal: a feasibility study. *Pediatr Radiol* 2022; **52**: 1104–14. https://doi.org/10.1007/s00247-021-05270-x

27. Wennmann M, Klein A, Bauer F, Chmelik J, Grözinger M, Uhlenbrock C, et al. Combining deep learning and radiomics for automated, objective, comprehensive bone marrow characterization from whole-body MRI: A multicentric feasibility study. *Invest Radiol* 2022. https://doi.org/10.1097/RLI.0000000000000891

28. G RA. Machine Learning in Myeloma Response (MALIMAR) study. Internet. 2022. Available from: https://clinicaltrials.gov/ct2/show/NCT03574454

29. Park HY, Kim KW, Yoon MA, Lee MH, Chae EJ, Lee JH, et al. Role of whole-body MRI for treatment response assessment in multiple myeloma: comparison between clinical response and imaging response. *Cancer Imaging* 2020; **20**: 14. https://doi.org/10.1186/s40644-020-0293-6

30. Horger M, Weisel K, Horger W, Mroue A, Fenchel M, Lichy M. Whole-body diffusion-weighted MRI with apparent diffusion coefficient mapping for early response monitoring in multiple myeloma: preliminary results. *AJR Am J Roentgenol* 2011; **196**: W790–95. https://doi.org/10.2214/AJR.10.5979

31. Wu C, Huang J, Xu W-B, Guan Y-J, Ling H-W, Mi J-Q, et al. Discriminating depth of response to therapy in multiple myeloma using whole-body diffusion-weighted MRI with apparent diffusion coefficient: preliminary results from a single-center study. *Acad Radiol* 2018; **25**: 904–14. https://doi.org/10.1016/j.acra.2017.12.008