








## Article

# Semi-Automated Segmentation of Bone Metastases from Whole-Body MRI: Reproducibility of Apparent Diffusion Coefficient Measurements

Alberto Colombo <sup>1,\*</sup>, Giulia Saia <sup>1</sup>, Alcide A. Azzena <sup>2</sup>, Alice Rossi <sup>3</sup>, Fabio Zugni <sup>1</sup>, Paola Pricolo <sup>1</sup>, Paul E. Summers <sup>1</sup>, Giulia Marvaso <sup>4,5</sup>, Robert Grimm <sup>6</sup>, Massimo Bellomi <sup>1,5</sup>, Barbara A. Jereczek-Fossa <sup>4,5</sup>, Anwar R. Padhani <sup>7</sup> and Giuseppe Petralia <sup>5,8</sup>

- <sup>1</sup> Division of Radiology, IEO European Institute of Oncology IRCCS, 20141 Milan, Italy; giulia.saia@ieo.it (G.S.); fabio.zugni@ieo.it (F.Z.); paola.pricolo@ieo.it (P.P.); paul.summers@ieo.it (P.E.S.); massimo.bellomi@ieo.it (M.B.)
  - <sup>2</sup> Postgraduate School in Radiodiagnostics, University of Milan, 20122 Milan, Italy; alcide.azzena@unimi.it
  - <sup>3</sup> Radiology Unit, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS, 47014 Meldola, Italy; alice.rossi@irst.emr.it
  - <sup>4</sup> Division of Radiotherapy, IEO European Institute of Oncology IRCCS, 20141 Milan, Italy; giulia.marvaso@ieo.it (G.M.); barbara.jereczek@ieo.it (B.A.J.-F.)
  - <sup>5</sup> Department of Oncology and Hemato-Oncology, University of Milan, 20122 Milan, Italy; giuseppe.petralia@ieo.it
  - <sup>6</sup> MR Applications Pre-Development, Siemens Healthcare, 91052 Erlangen, Germany; robertgrimm@siemens-healthineers.com
  - <sup>7</sup> Paul Strickland Scanner Centre, Mount Vernon Cancer Centre, Northwood HA6 2RN, UK; anwar.padhani@stricklandscanner.org.uk
  - <sup>8</sup> Precision Imaging and Research Unit, Department of Medical Imaging and Radiation Sciences, IEO European Institute of Oncology IRCCS, 20141 Milan, Italy
- \* Correspondence: alberto.colombo@ieo.it



**Citation:** Colombo, A.; Saia, G.; Azzena, A.A.; Rossi, A.; Zugni, F.; Pricolo, P.; Summers, P.E.; Marvaso, G.; Grimm, R.; Bellomi, M.; et al. Semi-Automated Segmentation of Bone Metastases from Whole-Body MRI: Reproducibility of Apparent Diffusion Coefficient Measurements. *Diagnostics* **2021**, *11*, 499. <https://doi.org/10.3390/diagnostics11030499>

Academic Editor: Evangelos Terpos

Received: 16 February 2021  
Accepted: 9 March 2021  
Published: 11 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Using semi-automated software simplifies quantitative analysis of the visible burden of disease on whole-body MRI diffusion-weighted images. To establish the intra- and inter-observer reproducibility of apparent diffusion coefficient (ADC) measures, we retrospectively analyzed data from 20 patients with bone metastases from breast (BCa;  $n = 10$ ; aged  $62.3 \pm 14.8$ ) or prostate cancer (PCa;  $n = 10$ ; aged  $67.4 \pm 9.0$ ) who had undergone examinations at two timepoints, before and after hormone-therapy. Four independent observers processed all images twice, first segmenting the entire skeleton on diffusion-weighted images, and then isolating bone metastases via ADC histogram thresholding (ADC:  $650\text{--}1400 \mu\text{m}^2/\text{s}$ ). Dice Similarity, Bland-Altman method, and Intraclass Correlation Coefficient were used to assess reproducibility. Inter-observer Dice similarity was moderate (0.71) for women with BCa and poor (0.40) for men with PCa. Nonetheless, the limits of agreement of the mean ADC were just  $\pm 6\%$  for women with BCa and  $\pm 10\%$  for men with PCa (mean ADCs: 941 and  $999 \mu\text{m}^2/\text{s}$ , respectively). Inter-observer Intraclass Correlation Coefficients of the ADC histogram parameters were consistently greater in women with BCa than in men with PCa. While scope remains for improving consistency of the volume segmented, the observer-dependent variability measured in this study was appropriate to distinguish the clinically meaningful changes of ADC observed in patients responding to therapy, as changes of at least 25% are of interest.

**Keywords:** WB-MRI; DWI; ADC; quantitative analysis; bone metastases; reproducibility

## 1. Introduction

Occurring in up to 70% of patients with advanced breast cancer (BCa) or prostate cancer (PCa), bone metastases are frequently present in patients in therapy for these tumours [1]. Precise and timely assessments of therapy response in metastatic BCa and PCa are needed to ensure targeted therapies are administered efficiently [2,3]. The RECIST v1.1

criteria commonly used for evaluating response to treatment, however, are inappropriate for assessing the response of bone metastases, because bone-limited lesions are classified as “unmeasurable” [4,5]. Whole-body MRI (WB-MRI) that includes WB diffusion-weighted images (DWI) marks a paradigm shift in the assessment of treatment response of bone metastases [6–8]: indeed, beyond volume changes visible on conventional imaging, WB-MRI can also detect early functional changes via differences in apparent diffusion coefficient (ADC) [9,10], a quantitative index of water motility obtained from DWI [11,12]. Unlike soft tissue lesions [13], active bone lesions have higher ADC values than normal, fat-rich bone marrow [14,15], but ADC values tend to increase for both soft tissues and malignant bone lesions when there is substantial response to therapy due to an increased mobility of water molecules accompanying cell death [16,17].

As a quantitative metric, each ADC measurement is subject to uncertainty related to patient and experimental variability (physiological factors, scanner used, DWI acquisition protocol and ADC computation method) as well as to the process of drawing regions of interest from which to extract the values [18]. This last process is particularly challenging in metastatic patients when multiple lesions are distributed in the skeleton. A recent systematic analysis reported that ADC differences of at least 12% in repeated experiments could be considered true changes [19], while the inter-observer variability of mean ADC in bone metastases was about 7% and thus sufficiently low not to significantly reduce overall sensitivity to clinically relevant ADC changes [20].

The potential of WB-MRI ADC histogram analysis for monitoring of bone disease has been demonstrated but its clinical use is limited because segmentation can be influenced by observer experience and is time-consuming to perform, even with semi-automation [20–24]. In response to these shortcomings, a streamlined semi-automatic technique for segmenting distributed bone metastases has been developed that combines the optional calculation of heavily diffusion-weighted images [25], with thresholds over the entire image volume, manual editing, and finally, limitation of ADC values to the range of clinical interest.

The aim of this study was to determine the intra- and inter-observer reproducibility for quantitative ADC values obtained through this semi-automated approach to segmentation of bone metastases from WB-MRI by multiple observers with widely varying prior experience.

## 2. Materials and Methods

### 2.1. Population

The local ethics committee approved this retrospective single center study, and written informed consent was obtained from the subjects for use of their data. Based on a power analysis using the results of a previous study [20], 20 patients with two WB-MRI examinations were included in the study. In order to obtain a homogeneous population from the point of view of the therapy performed, patients were consecutively included if compliant with these criteria: having undergone both a baseline WB-MRI examination prior to initiating therapy and a follow-up WB-MRI examination during first-line hormone therapy following a radiological diagnosis of metastatic bone disease originating for women with invasive ductal or lobular breast carcinoma and for men with prostatic adenocarcinoma between January 2013 and March 2018. Both examinations were included in the study to represent the range of examinations occurring in clinical routine. Patients who underwent other metastases directed treatments (chemotherapy, radiotherapy, surgery) before the follow-up study were excluded.

## 2.2. WB-MRI Acquisition Protocol

The WB-MRI examinations were performed using a 1.5T MR scanner (MAGNETOM Avanto<sup>fit</sup>, Siemens Healthcare, Erlangen, Germany). The scanning protocol was MET-RADS-P compliant [6]. In particular, DWI scans extended from the upper border of the orbits to mid-thigh and consisted of four contiguous stations of 50 slices acquired in free-breathing using a 2D single shot echo-planar imaging (SS-EPI) sequence. Over the course of the study, two distinct shimming techniques and acquisition parameter sets (Table 1) were used for the DWI scans without changes to the  $b$ -values applied. Initially, a single, volumetric shim was determined for each station and applied for all slices within the station. From June 2016 onwards, slice-specific shimming was performed within each station using a prototype acquisition software provided by the machine vendor [26].

**Table 1.** Whole-body diffusion-weighted image acquisition protocols.

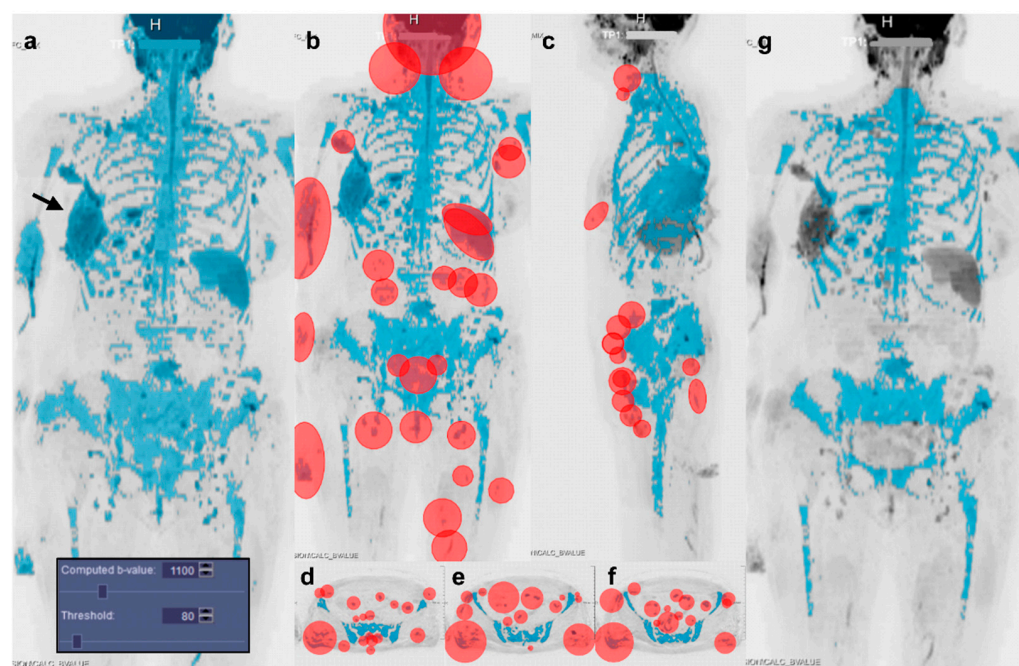
	Station-Specific Shim	Slice-Specific Shim
Sequence	Diffusion-Weighted SS-EPI	Diffusion-Weighted SS-EPI
Orientation	Transversal	Transversal
$b$ -value (s/mm <sup>2</sup> )	50, 900	50, 900
Encoding mode	3-scan trace	3D-diagonal
Averages per $b$ -value	6, 6	5, 15
Repetition Time (ms)	9000	6550
Echo Time (ms)	67	62
Fat Saturation	STIR	STIR
Inversion Time (ms)	180	180
Field of View (mm)	337 × 450	390 × 429
Slice thickness (mm)	5.0	5.0
Gap between slices (mm)	0.0	0.0
Voxel size (mm <sup>3</sup> )	1.8 × 1.8 × 5.0	1.6 × 1.6 × 5.0
Acquisition time (min)	22:00	15:02

SS-EPI = Single-Shot Echo-Planar Imaging, STIR = Short Tau Inversion Recovery.

## 2.3. Image Segmentation

The WB-MRI examination images were exported in DICOM format to a reporting workstation. To allow independent evaluation of the baseline and follow-up examinations, a distinct code was applied to each examination during anonymization. Segmentation of bone metastases consisted of two sub-procedures.

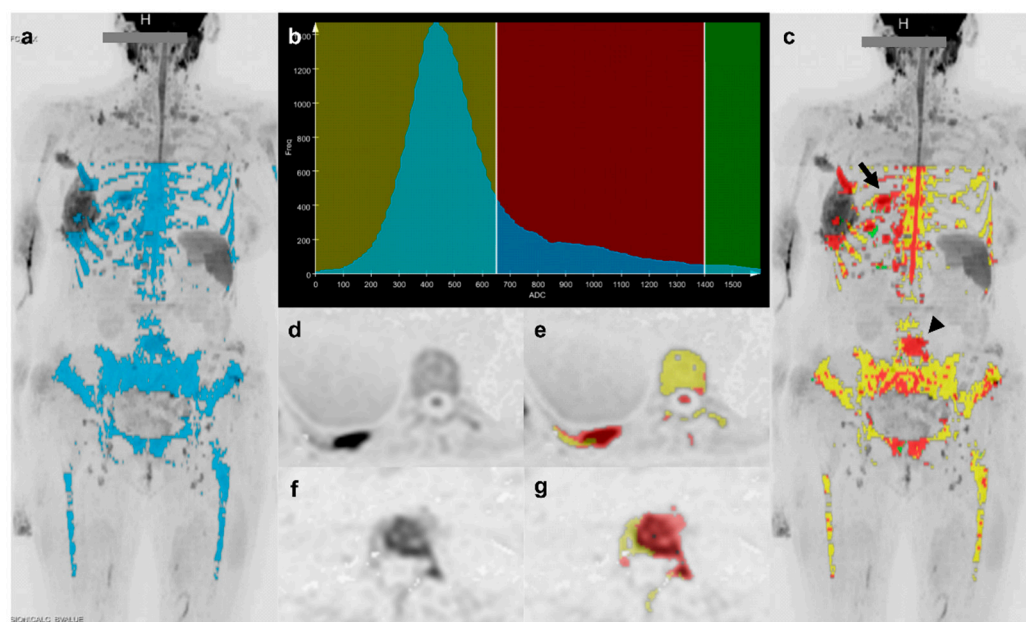
First, bone marrow segmentation was performed using a semi-automatic approach based on signal intensity thresholding of DWI images, previously described for the direct segmentation of the visible bone metastases [22,25]. A software implementation of this method (MR Total Tumor Load, Siemens Healthcare, Erlangen, Germany) was used, that combined automatic pre-processing and computation of a volumetric ADC map (with mono-exponential fitting); observers were required to select a signal intensity threshold applied to a simulated high  $b$ -value image stack and manual editing to obtain a bone marrow mask (Figure 1).



**Figure 1.** Illustration of the semi-automatic method used for bone marrow segmentation. Bone segmentation started with the observer interactively selecting a  $b$ -value for the calculation of a diffusion-weighted image stack (computed  $b$ -value) that provided good visual contrast between bone and surrounding tissues in coronal inverted gray-scale maximum intensity projection (MIP). (a) The observer then interactively adjusted a threshold to isolate voxels having high signal intensity on the computed  $b$ -value image (i.e., darker on the inverted MIP) to incorporate as much bone as possible in the resulting mask (seen overlaid in blue on the inverted MIP of the computed  $b$ -value DWI stack). The initial mask thus included suspected hypercellular lesions and as much bone as possible, but inevitably also included some non-bone tissues, typically brain and spinal cord, spleen, male gonads, breast implants, and sites of soft tissue inflammation or soft tissue lesions (e.g., the large soft tissue metastases along the right chest wall (arrow) and supraclavicular lymph nodes seen in this case of a 71-year-old woman with operated lobular carcinoma of the right breast, undergoing endocrine treatment). Manual editing was therefore performed to remove as much non-bone tissue as possible using a combination of: (b) full-depth cutting of ellipsoids (overlaid in red) positioned on the coronal MIP to eliminate brain, neck lymph nodes, soft tissues of the small pelvis, and as needed, spleen, kidneys, and lymph nodes not overlapping diseased bone, followed by (c) full-depth cutting of ellipsoids (overlaid in red) drawn on the sagittal MIP to eliminate soft tissues of the anterior neck, breast implants (if any), bowel, rectum, as well as inguinal and external iliac lymph nodes. If needed, (d,e,f) single-slice cutting of ellipsoids (overlaid in red) on individual axial, coronal or sagittal slices to eliminate soft tissues that projected over bone in the MIPs. Finally, the bone mask (g) was saved as a DICOM image stack.

Second, the bone mask and the ADC map were saved as DICOM image stacks and used in calculating ADC histograms to isolate the metastases via ad hoc functions written in Python 3.7 (Python Software Foundation, Beaverton, OR, USA). In short, a lower threshold of  $650 \mu\text{m}^2/\text{s}$  and an upper threshold of  $1400 \mu\text{m}^2/\text{s}$  were applied to the masked regions of the ADC map to remove normal bone marrow [27–30] and necrotic disease [6] (Figure 2). For the remaining voxels, which were assumed to represent bone lesions, we calculated the segmentation volume (Volume), mean (Mean\_ADC), standard deviation (Std\_ADC), median (Median\_ADC), 5th and 95th percentiles (5%\_ADC and 95%\_ADC), skewness (Skewness\_ADC), kurtosis (Kurtosis\_ADC), and histogram entropy (Entropy\_ADC) from ADC histograms. Due to signal differences between the stations obtained with the head/neck coil and the remaining body stations (acquired with anterior and posterior array coils),

we limited our processing to the three body blocks covering from the upper thorax to the mid-thighs.



**Figure 2.** Extracting the bone metastases region of interest by applying apparent diffusion coefficient (ADC) thresholds to the bone marrow mask. Lesion segmentation started with the bone mask (a) seen overlaid on a coronal inverted gray-scale maximum intensity correction (MIP) being used to produce an ADC histogram (b) from the ADC data. The histogram was divided into three categories on the basis of two thresholds: below  $650 \mu\text{m}^2/\text{s}$  corresponding to normal bone (yellow band), between  $650 \mu\text{m}^2/\text{s}$  and  $1400 \mu\text{m}^2/\text{s}$  corresponding to lesions (red band), and above  $1400 \mu\text{m}^2/\text{s}$  corresponding to necrotic lesion or cyst (green band). The normal bone and bone lesion voxels identified in this way were then colored as yellow and red, respectively, and overlaid on the coronal inverted gray-scale MIP (c) to show the localization of healthy bone and active disease. For two lesions (arrow and arrowhead in (c)), axial high  $b$ -value diffusion-weighted images and the resulting separation of bone marrow (in yellow) from metastases (in red) shows: a lesion of the posterior arc of the 10th right rib (arrow in (c–e)); a lesion of the lumbosacral spine involving transverse process and part of the vertebral body (arrowhead in (c,f,g)). Some residual soft tissues having ADC values in the considered range were included in the final evaluation (e.g., spinal cord in (e)).

#### 2.4. Observers

Four independent observers segmented each of the 40 DWI scans and repeated the process at least three weeks later, in a separate reading session, to minimize recall bias. None of the observers had prior experience in reporting WB-MRI. Two observers, a biomedical engineer experienced in image processing (Obs1\_MASKED) and a radiologist with eight years of experience (Obs3\_MASKED) had relevant background expertise in medical image processing, while the other two—a radiology resident (Obs2\_MASKED) and a student radiology technologist (Obs4\_MASKED)—were relatively inexperienced in image processing methods.

#### 2.5. Statistical Analysis

The computed  $b$ -values and thresholds chosen by the observers for segmentation, and the time required to complete each segmentation were recorded. The similarity between segmentations was expressed using the Dice Similarity Coefficient (DSC) [31]. Associations between DSC and factors potentially influencing segmentation similarity were assessed using factorial ANOVA and Spearman's coefficient ( $\rho_s$ ) for categorical and continuous variables, respectively. The factors considered were patient sex, age, treatment status at the time of WB-MRI (baseline or follow-up examination), number of MET-RADS-P skeletal

regions with metastases, and shimming technique [6]. In light of a strong effect of sex, subsequent analyses were performed separately for men with PCa and women with BCa, and the Mann-Whitney test was used to compare measures.

The distribution of Mean\_ADC and of the other histogram parameter values was expressed as their average across readers. Measures of the second reading session were considered in order to minimize the learning curve effect for the first segmentation.

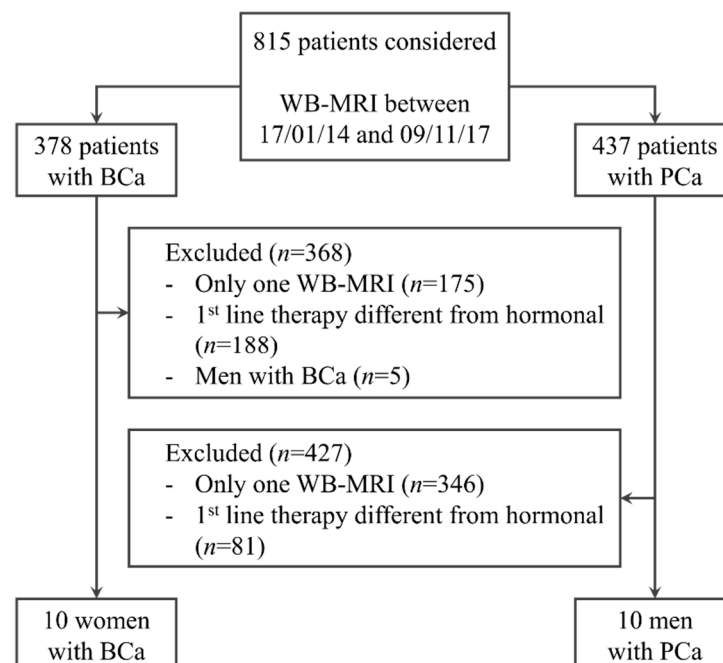
The Bland-Altman method [32,33] was used to evaluate intra-observer (comparing the first and second segmentations for each observer), and inter-observer (comparing pairs of readers within each of the two reading sessions) reproducibility of Mean\_ADC. Dependence of absolute differences on the mean of measurements was assessed using Kendall's tau ( $\tau_b$ ), and the mean intra- and inter-observer bias and 95% limits of agreement were determined. The correlation between the volume of segmentation and the variability of Mean\_ADC (mean difference among intra-observer and inter-observer measurements) was evaluated with Spearman's correlation coefficient.

Intra- and inter-observer reproducibility of the other ADC histogram statistics were measured using Intraclass Correlation Coefficients (ICC) with 95% confidence intervals, calculating absolute concordance using a two-way model with mixed effects and single measurements [34]. For both DSC and ICC, the following classification scale was used to evaluate similarity/reproducibility: poor ( $DSC/ICC < 0.50$ ), modest ( $0.50 \leq DSC/ICC < 0.75$ ), good ( $0.75 \leq DSC/ICC < 0.90$ ), and excellent ( $DSC/ICC \geq 0.90$ ). We considered results of  $p < 0.05$  significant and analyses were performed with the R software package (R 2018, version 3.5.1, Vienna, Austria).

### 3. Results

#### 3.1. Population

Of the 378 BCa and 437 PCa patients who underwent WB-MRI in the study period (Figure 3), 10 women with BCa and 10 men with PCa met the inclusion criteria. Clinical and demographic characteristics of the patients are summarized in Table 2.



**Figure 3.** Diagram of the patient selection workflow. Of the 815 patients who had undergone whole-body MRI (WB-MRI) in our institution during the study period, 20 patients, 10 men with prostate cancer (PCa) and 10 women with breast cancer (BCa), satisfied the inclusion criteria of having undergone more than one WB-MRI, and being on first-line hormonal therapy.

**Table 2.** A summary of clinical and demographical information.

		Men with PCa	Women with BCa
Patients	Number	10	10
	Age at baseline (years) <sup>1</sup>	67.4 (51–77)	62.3 (31–76)
	No. of skeletal regions with metastasis <sup>1</sup>	3.7 (2–6)	5.0 (3–6)
Type of primary tumour	Prostatic adenocarcinoma	10	0
	Invasive ductal breast cancer	0	3
	Invasive lobular breast cancer	0	7
No. of patients with metastasis for each skeletal region	Skull	0	2
	Spine (cervical)	3	9
	Spine (thorax)	6	7
	Spine (lumbosacral)	8	9
	Thorax	7	8
	Pelvis	9	10
	Limbs	4	7
Other sites of disease	Lymph nodes	60% (6/10)	50% (5/10)
	Visceral	0% (0/10)	40% (4/10)
	Local disease	50% (5/10)	10% (1/10)
	Other (muscles)	0% (0/10)	10% (1/10)
WB-MRI examinations	Observation period	17/01/14–29/05/17	03/09/15–09/11/17
	No. baseline/follow-up	10/10	10/10
	Days between baseline and follow-up <sup>1</sup>	213.8 (90–373)	197.9 (109–291)
	No. with station-/slice-specific shim	8/12	15/5

WB-MRI = Whole-Body MRI, PCa = Prostate cancer, BCa = Breast cancer, <sup>1</sup> Mean (range).

Follow-up examinations were performed an average 206 days after the baseline examination (range: 90–373 days). Of the 40 WB-MRI examinations analyzed, 23 were acquired using station-based shimming (15 BCa and 8 PCa patients), and 17 with slice-specific shimming (5 BCa and 12 PCa patients).

### 3.2. Segmentation Settings and Duration

The time between the first and the second reading sessions ranged from three to four weeks across the four observers. A summary of the settings used, and times required for evaluation is given in Table S1 (Supplementary Material). Between the observers, the average  $b$ -values used for the computed  $b$ -value image ranged from  $994 \pm 23$  to  $1057 \pm 67$  s/mm<sup>2</sup>, with an overall mean of 1012 s/mm<sup>2</sup>. The threshold signal intensity for initial segmentation ranged from  $32.5 \pm 20.5$  to  $47.3 \pm 41.6$  with an overall mean of 41.

The time required to perform a segmentation ranged from 4 to 38 min. For the experienced observers (Obs 1, Obs 3), the average segmentation time across all examinations was about 11 min shorter (12 vs. 23 min) and the range of times for individual patients narrower (4 to 28 min vs. 8 to 38 min) than for the inexperienced observers (Obs 2, Obs 4). On average, the observers were  $2.1 \pm 0.4$  min faster in the second segmentation session.

### 3.3. Factors Influencing Segmentation Similarity

Patient sex was significantly associated with mean intra-observer DSC values ( $p < 0.0001$ ), which were greater for women with BCa (Figure S1, Supplementary Material). A smaller, but still significant association was also seen with respect to the shimming technique used ( $p < 0.01$ ), with station-based shimming tending to yield a higher DSC. Treatment (baseline vs. follow-up examination) had no effect ( $p = 0.81$ ). A moderate positive correlation between DSC and number of skeletal regions with metastases ( $\rho_s = 0.58$ ,  $p < 0.0001$ ) was also noted.

### 3.4. Distribution of Quantitative Parameters Values

In women with BCa, the average Mean\_ADC measurement of the four observers was  $936.6 \pm 101.9 \mu\text{m}^2/\text{s}$  at baseline, and  $945.4 \pm 91.3 \mu\text{m}^2/\text{s}$  at follow-up WB-MRI. Similar values were found in men with PCa, for whom Mean\_ADC was  $963.5 \pm 91.5 \mu\text{m}^2/\text{s}$  and  $1033.5 \pm 84.1 \mu\text{m}^2/\text{s}$  at baseline and follow-up, respectively. The Table S2 (Supplementary Material) shows the distribution of average values, at baseline and follow-up, for the other quantitative histogram parameters.

### 3.5. Intra- and Inter-Observer Reproducibility Analysis

Overall, the mean intra-observer DSC value was modest (0.67) but was significantly higher in women with BCa than in men with PCa (good: 0.78 vs. modest: 0.55,  $p < 0.0001$ ).

For women with BCa, the intra-observer Bland-Altman bias and limits of agreement of Mean\_ADC were 0.5% (−5.2%, 6.0%) for an average measure of  $942.9 \mu\text{m}^2/\text{s}$ , and for men with PCa, they were 0.5% (−9.0%, 9.9%) and  $1000.8 \mu\text{m}^2/\text{s}$ . No significant correlation was found between the volume of lesion segmented and the variability of Mean\_ADC in women with BCa ( $\rho_s = -0.35$ ,  $p = 0.13$ ), or in men with PCa ( $\rho_s = 0.15$ ,  $p = 0.50$ ). In women with BCa, the intra-observer ICC for Mean\_ADC showed excellent agreement (95% CI, 0.90–0.98), while in men with PCa, it was modest to excellent (95% CI, 0.63–0.92). Across the parameters considered, the intra-observer ICCs tended to be greater, and the 95% confidence intervals narrower, for women with BCa than for men with PCa.

Results of the intra- and inter-observer reproducibility analyses are summarized in Table 3. Detailed information regarding DSC, Bland Altman bias and limits of agreement and ICC are reported, respectively, in Tables S3–S5 (Supplementary Material). Neither intra- nor inter-observer differences in ADC showed relevant dependence on mean ADC ( $\tau_b = 0.12$  with  $p = 0.27$  and  $\tau_b = -0.34$  with  $p < 0.01$ ).

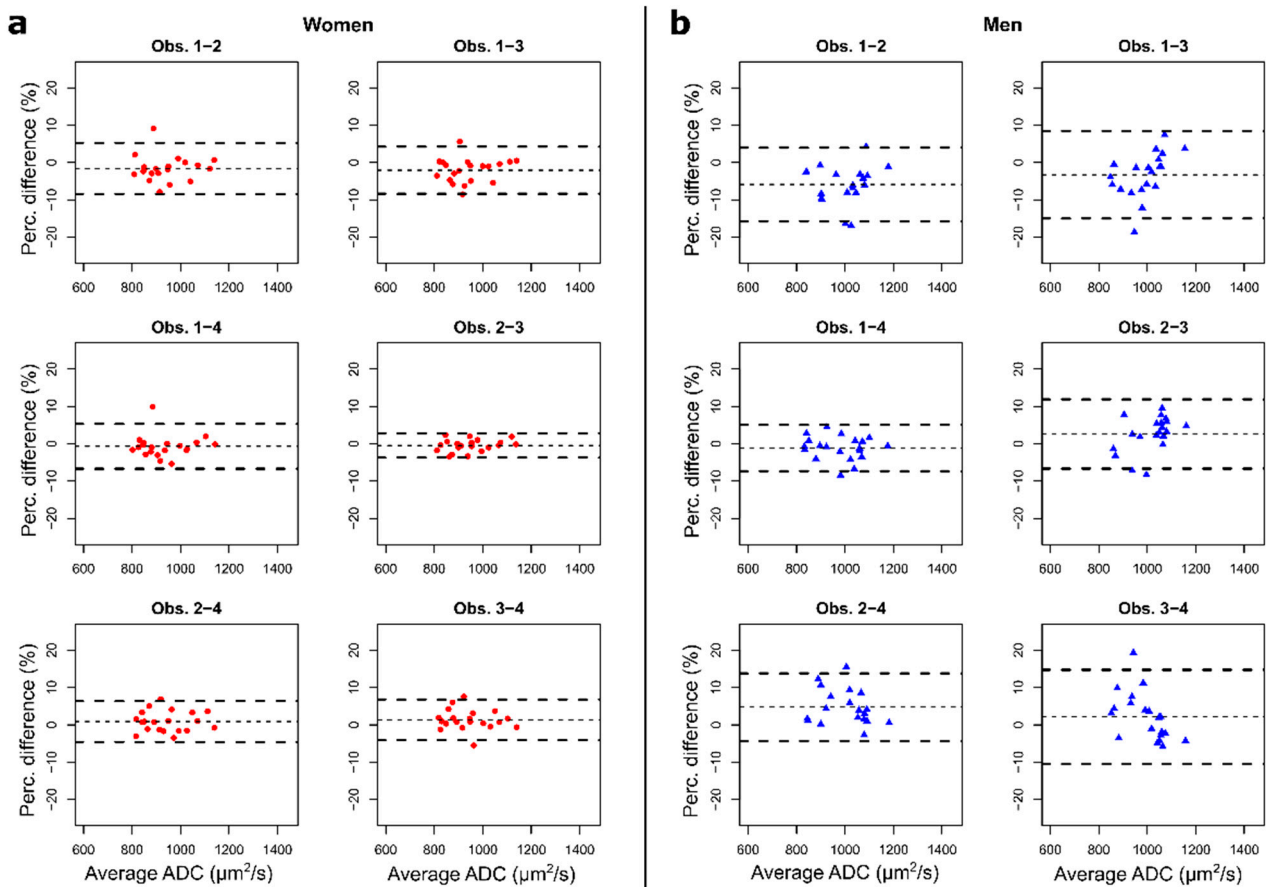
**Table 3.** Summary of observer-dependent reproducibility metrics.

		Women with BCa ( <i>n</i> = 20)	Men with PCa ( <i>n</i> = 20)
<b>Intra-observer</b>			
DSC	mean $\pm$ std dev	0.78 $\pm$ 0.14	0.55 $\pm$ 0.21
BA (Mean_ADC)	bias (LoA)	0.5% (−5.2%, 6.0%)	0.5% (−9.0%, 9.9%)
ICC (Mean_ADC)	estimate (95% CI)	0.96 (0.90–0.98)	0.82 (0.63–0.92)
<b>Inter-observer (2nd reading)</b>			
DSC	mean $\pm$ std dev	0.71 $\pm$ 0.16	0.40 $\pm$ 0.19
BA (Mean_ADC)	bias (LoA)	1.2% (−6.0%, 5.1%)	3.3% (−9.9%, 9.6%)
ICC (Mean_ADC)	estimate (95% CI)	0.96 (0.91–0.98)	0.79 (0.60–0.91)

BCa = breast cancer, PCa = prostate cancer, DSC = Dice Similarity Coefficient, BA = Bland-Altman, LoA = Limits of Agreement, ICC = Intraclass Correlation Coefficients, Mean\_ADC = mean of apparent diffusion coefficient distribution.

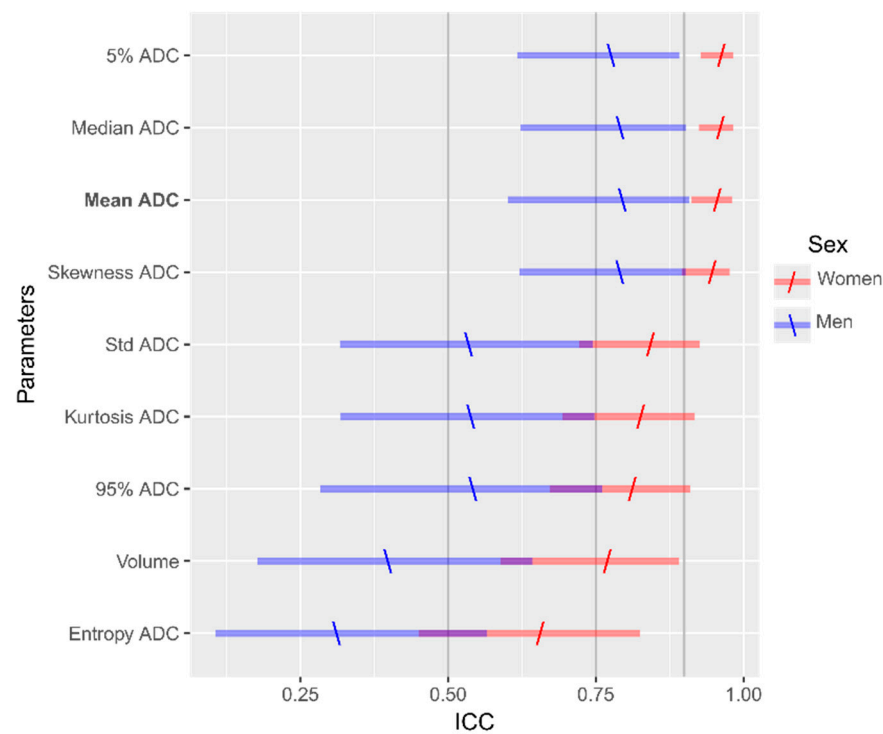
Overall, the mean inter-observer DSC showed modest segmentation similarity (0.52 and 0.55 for first and second reading respectively). The DSC values were significantly higher in women with BCa (0.67 for the first reading and 0.71 for the second) than in men with PCa ( $p < 0.0001$ ), where poor similarity was observed (0.37 and 0.40). Inter-observer bias and limits of agreement of Mean\_ADC in the second reading session for women with BCa and men with PCa were, respectively, 1.2% (−6.0%, 5.1%) and 3.3% (−9.9%, 9.6%), for average measures of  $941.0 \mu\text{m}^2/\text{s}$  and  $998.5 \mu\text{m}^2/\text{s}$  (Figure 4).





**Figure 4.** Bland-Altman plots of inter-observer mean apparent diffusion coefficient (ADC) measures of bone lesions. Each plot represents the percentage difference between the measures of a pair of observers compared to the average of their measures in the second reading session. In our cohort, (a) excellent reproducibility was observed in women with breast cancer, with bias and 95% limits of agreement below  $\pm 2.5\%$  and  $\pm 8.5\%$ , respectively. (b) Higher variability was observed in men with prostate cancer, with bias and 95% limits of agreement below  $\pm 6\%$  and  $\pm 16\%$ , respectively.

In the inter-observer analysis, the lesion volume and variability of Mean\_ADC were not correlated: though a weak negative trend was observed in women with BCa ( $\rho_s = -0.42$ ,  $p = 0.07$ ), it was not seen in men with PCa ( $\rho_s = -0.16$ ,  $p = 0.49$ ). The inter-observer ICC for Mean\_ADC showed excellent reproducibility in women with BCa (95% CI, 0.91–0.98), as opposed to the modest to excellent reproducibility obtained in men with PCa (95% CI, 0.60–0.91). The inter-observer ICC analysis of the other histogram statistics showed greater reproducibility and narrower 95% confidence intervals, for women with BCa than for men with PCa (Figure 5).



**Figure 5.** Graphical representation of the inter-observer Intraclass Correlation Coefficients (ICC) with lower and upper limits of the 95% confidence intervals calculated for parameters derived from the apparent diffusion coefficient histogram. The population is divided by sex (blue: men with prostate cancer, red: women with breast cancer), back-slashes and forward slashes represent the estimated ICC values for men and women, respectively. For our cohort, the intervals were narrower and ICC values nearer to 1 in women, indicating greater reproducibility than in men.

#### 4. Discussion

Prior studies have demonstrated the ability of WB-MRI-based ADC measurements to monitor treatment response in patients with metastatic bone disease [20]. However, available approaches to segmentation of bone metastases are dependent on radiological expertise and are too time consuming for realistic clinical use [22]. As a precursor to the use of WB-MRI in the monitoring of treatment in metastatic disease, we examined the intra- and inter-observer reproducibility of metastatic bone lesion segmentation and of the corresponding ADC values obtained using a semi-automated tool for segmenting dispersed skeletal lesions, by observers with diverse clinical expertise.

Our process starts with segmentation of bone, for which the  $b$ -values chosen to provide optimal contrast between bone marrow and soft tissues on the calculated diffusion weighted image were close to  $1000 \text{ s/mm}^2$  across the cohort of patients. Blackledge et al. [25] found similar  $b$ -values yielded simulated images (median:  $1070 \text{ s/mm}^2$ , range:  $715\text{--}1660 \text{ s/mm}^2$ ) that were optimal for direct lesion segmentation. Both results point to the optimal  $b$ -value for segmentation being different from that recommended for acquisition in the MET-RADSP and MY-RADS guidelines ( $800 \text{ s/mm}^2$ ) where scan time and contrast to noise must be accommodated [6,35]. This is not a significant obstacle as the calculation of a higher  $b$ -value image for use in segmentation can readily be obtained via a mono-exponential calculation.

The second step in the segmentation process involved the selection of a threshold based on the calculated  $b$ -value image. Due to a lack of standardization of the MRI signal intensities, this threshold is likely to depend on acquisition settings, field-strength, and system hardware. Normalization of the DWI signal intensities by the muscle signal intensity has been proposed as a strategy for threshold selection that is independent of the diffusion MRI acquisition settings (e.g., gain settings,  $b$ -value gradient, coil and fat-suppression method) [27].

The two more experienced observers took an average of 12 min (range 4–28 min) to complete each segmentation, while the less experienced observers averaged 23 min. These times compare favourably with the roughly 30 min reported by Blackledge et al. [22], for segmentation of metastases by experienced observers. Achieving segmentation in a clinically acceptable time is a key obstacle to be overcome for the use of quantitative WB-MRI in monitoring treatment response in bone metastases.

In assessing factors that influence segmentation similarity, as indicated by the mean intra-observer DSC values, we found that patient sex had a particularly strong effect, with higher DSC values being seen in women with BCa than in men with PCa (0.78 vs. 0.55). We attribute this difference to hyperintensity of bone marrow on diffusion-weighted images in women, a feature observed in previous studies [36,37]. If the marrow is hyperintense, the threshold applied to the high *b*-value images allows a cleaner separation from other tissues, and thus requires less manual editing. This makes semi-automatic segmentation of bone marrow particularly suitable for women, while additional post-processing would be required in men to achieve matching levels of segmentation similarity [38]. Shimming technique had a small but significant effect on DSC values. This likely relates to different signal-to-noise ratios in the diffusion-weighted images due to the difference in shim quality, but has not been found to result in significant differences in ADC values [26,39]. We have therefore incorporated data obtained using both shim techniques in this study to evaluate the variability related to the observers performing the analysis on each image independently.

Taking the inter-observer 95% limits of agreement of Mean\_ADC in bone metastases as representative of observer performance, changes of 6% in women with BCa and 10% in men with PCa could be considered beyond the observer-related variability. These values are similar to the 7% reported by Blackledge et al. [20], who used a more time-consuming approach to segmenting bone metastases. These differences suggest that, with this method, a better sensitivity to ADC change in metastases can be expected for women with BCa than for men with PCa.

On top of the observer-related variability documented here, test-retest experimental variability needs to be considered to establish the magnitude of change in ADC that must occur before it can be unequivocally detected. Winfield et al. reported that mean ADC increases of >12% could be considered real changes in repeated experiments [19]. It is reasonable therefore, to expect that the test-retest variability of WB-MRI for ADC of bone metastases will be clinically acceptable as the MET-RADS-P and MY-RADS guidelines indicate that increases in Mean\_ADC values induced by therapy should be at least 25% between baseline and follow-up in case of “likely” response, and at least 40% in case of “very likely” response [6,35].

Blackledge et al. [20] reported excellent reproducibility not only for mean ADC, but also for metastatic volume parameters, consistent with the results of Perez-Lopez et al. [23]. While our results were similar for reproducibility of mean and median ADC, reproducibility of the segmentation volume was lower, yielding poor-modest DSC values. The variability of the lesion volumes and of the DSC values on the other hand, likely reflects the remaining subjectivity in the initial segmentation and in the manual elimination of residual soft tissues. This variability limits the clinical applicability of volume-related measures obtained with this method, particularly for evaluating men with PCa. Future studies should seek to reduce the variability of volume among readers while improving the repeatability of disease segmentation. The relative immunity of the ADC values in the face of variable lesion volume, suggests that the use of the thresholds for isolating the metastases imposes a degree of robustness in terms of ADC values, making this parameter an interesting complementary tool to radiological evaluation.

The small sample size and single center nature of our study are limitations that may restrict generalizability to our work. In addition, there was no ground truth for the segmentation of metastases (e.g., manual segmentation performed by a radiologist experienced in WB-MRI), and consequently we cannot comment on segmentation accuracy. Furthermore, we have addressed only the issues of intra and inter-observer reproducibility.

Having found robust ADC value extraction, it is reasonable to pursue a test-retest study to establish the magnitude of change that can be detected with confidence. Finally, two different shimming techniques (station- and slice-specific) were used during the observation period: other studies have found the difference in ADC values between these techniques to be small, but their inclusion may have inflated the observer-related variability.

## 5. Conclusions

While scope remains for improving the consistency of the volume of bone metastases segmented, the segmentation method evaluated in this study demonstrates good to excellent levels of intra- and inter-observer reproducibility in measuring mean ADC, particularly for women with BCa. Noting that, according to MET-RADS-P and MY-RADS guidelines, the cut-off for clinically meaningful changes in mean ADC in patients who respond to therapy is at least 25%, the observer-dependent variability with the proposed approach is acceptable. Although observer-dependent variability was greater in men with PCa, the technique is likely to still be adequate for detecting responses to therapy at higher mean ADC change thresholds.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/2075-4418/11/3/499/s1>, Figure S1: Effect of sex, treatment status, and shimming technique on Dice Similarity Coefficients (DSC), Table S1: Segmentation settings and duration, Table S2: Distribution descriptors of parameters measured in the second reading, Table S3: Summary tables of intra-observer (on-diagonal) and inter-observer (off-diagonal) mean Dice Similarity Coefficients (DSC) by reading, Table S4: Summary of intra- and inter-observer Bland-Altman analysis results (bias and limits of agreement) for mean apparent diffusion coefficient (ADC) as percentages of the mean of the measures, Table S5 Intra- and inter-observer Intra-class Correlation Coefficients (ICC).

**Author Contributions:** Conceptualization, A.C., P.E.S. and G.P.; Data curation, A.C., G.S., A.A.A. and F.Z.; Investigation, A.C., G.S., A.A.A. and A.R.; Supervision, F.Z., P.P., P.E.S. and G.P.; Writing: original draft, A.C.; Writing: review & editing, P.P., P.E.S., G.M., R.G., M.B., B.A.J.-F. and A.R.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by FIEO-CCM. It was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds.

**Institutional Review Board Statement:** This retrospective study was approved by the Ethics Committee of European Institute of Oncology (R952/19-IEO 999, 03/04/2019).

**Informed Consent Statement:** Written informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data is available for review from the corresponding author on request.

**Acknowledgments:** The authors thank T. Benkert and F. Carpanese (Siemens Healthcare, Erlangen Germany) for the development of tools used in this study and the support.

**Conflicts of Interest:** The authors have no conflict of interest to declare. Exclusively Robert Grimm is an employee of Siemens Healthcare.

## References

1. Roodman, G.D. Mechanisms of Bone Metastasis. *N. Engl. J. Med.* **2004**, *350*, 1655–1664. [[CrossRef](#)]
2. Padhani, A.R.; Gogbashian, A. Bony metastases: Assessing response to therapy with whole-body diffusion MRI. *Cancer Imaging* **2011**, *11*, S129–S154. [[CrossRef](#)] [[PubMed](#)]
3. Zugni, F.; Ruju, F.; Pricolo, P.; Alessi, S.; Iorfida, M.; Colleoni, M.A.; Bellomi, M.; Petralia, G. The added value of whole-body magnetic resonance imaging in the management of patients with advanced breast cancer. *PLoS ONE* **2018**, *13*, e0205251. [[CrossRef](#)] [[PubMed](#)]
4. Eisenhauer, E.A.; Therasse, P.; Bogaerts, J.; Schwartz, L.H.; Sargent, D.; Ford, R.; Dancey, J.; Arbuck, S.; Gwyther, S.; Mooney, M.; et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **2009**, *45*, 228–247. [[CrossRef](#)] [[PubMed](#)]
5. Lecouvet, F.E.; Larbi, A.; Pasoglou, V.; Omoumi, P.; Tombal, B.; Michoux, N.; Malghem, J.; Lhommel, R.; Vande Berg, B.C. MRI for response assessment in metastatic bone disease. *Eur. Radiol.* **2013**, *23*, 1986–1997. [[CrossRef](#)]

6. Padhani, A.R.; Lecouvet, F.E.; Tunariu, N.; Koh, D.-M.; De Keyzer, F.; Collins, D.J.; Sala, E.; Schlemmer, H.P.; Petralia, G.; Vargas, H.A.; et al. METastasis Reporting and Data System for Prostate Cancer: Practical Guidelines for Acquisition, Interpretation, and Reporting of Whole-body Magnetic Resonance Imaging-based Evaluations of Multiorgan Involvement in Advanced Prostate Cancer. *Eur. Urol.* **2017**, *71*, 81–92. [[CrossRef](#)]
7. Petralia, G.; Padhani, A.R. Whole-Body Magnetic Resonance Imaging in Oncology: Uses and Indications. *Magn. Reson. Imaging Clin. N. Am.* **2018**, *26*, 495–507. [[CrossRef](#)]
8. Pricolo, P.; Ancona, E.; Summers, P.; Abreu-Gomez, J.; Alessi, S.; Jereczek-Fossa, B.A.; De Cobelli, O.; Nolè, F.; Renne, G.; Bellomi, M.; et al. Whole-body magnetic resonance imaging (WB-MRI) reporting with the METastasis Reporting and Data System for Prostate Cancer (MET-RADS-P): Inter-observer agreement between readers of different expertise levels. *Cancer Imaging* **2020**, *20*, 77. [[CrossRef](#)]
9. Pickles, M.D.; Gibbs, P.; Lowry, M.; Turnbull, L.W. Diffusion changes precede size reduction in neoadjuvant treatment of breast cancer. *Magn. Reson. Imaging* **2006**, *24*, 843–847. [[CrossRef](#)]
10. Galbán, C.J.; Hoff, B.A.; Chenevert, T.L.; Ross, B.D. Diffusion MRI in early cancer therapeutic response assessment. *NMR Biomed.* **2017**, *30*, e3458. [[CrossRef](#)]
11. Le Bihan, D. Apparent Diffusion Coefficient and Beyond: What Diffusion MR Imaging Can Tell Us about Tissue Structure. *Radiology* **2013**, *268*, 318–322. [[CrossRef](#)] [[PubMed](#)]
12. Li, S.P.; Padhani, A.R. Tumor response assessments with diffusion and perfusion MRI. *J. Magn. Reson. Imaging* **2012**, *35*, 745–763. [[CrossRef](#)] [[PubMed](#)]
13. Ahlawat, S.; Fayad, L.M. Diffusion weighted imaging demystified: The technique and potential clinical applications for soft tissue imaging. *Skelet. Radiol.* **2018**, *47*, 313–328. [[CrossRef](#)] [[PubMed](#)]
14. Dietrich, O.; Geith, T.; Reiser, M.F.; Baur-Melnyk, A. Diffusion imaging of the vertebral bone marrow. *NMR Biomed.* **2017**, *30*, e3333. [[CrossRef](#)] [[PubMed](#)]
15. Park, G.E.; Jee, W.-H.; Lee, S.-Y.; Sung, J.-K.; Jung, J.-Y.; Grimm, R.; Son, Y.; Paek, M.Y.; Min, C.-K.; Ha, K.-Y. Differentiation of multiple myeloma and metastases: Use of axial diffusion-weighted MR imaging in addition to standard MR imaging at 3T. *PLoS ONE* **2018**, *13*, e0208860. [[CrossRef](#)] [[PubMed](#)]
16. Padhani, A.R.; Koh, D.M.; Collins, D.J. Whole-body diffusion-weighted MR imaging in cancer: Current status and research directions. *Radiology* **2011**, *261*, 700–718. [[CrossRef](#)]
17. Petralia, G.; Padhani, A.R.; Pricolo, P.; Zugni, F.; Martinetti, M.; Summers, P.E.; Grazioli, L.; Colagrande, S.; Giovagnoni, A.; Bellomi, M. Whole-body magnetic resonance imaging (WB-MRI) in oncology: Recommendations and key uses. *Radiol. Med.* **2019**, *124*, 218–233. [[CrossRef](#)]
18. Barnes, A.; Alonzi, R.; Blackledge, M.; Charles-Edwards, G.; Collins, D.J.; Cook, G.; Coutts, G.; Goh, V.; Graves, M.; Kelly, C.; et al. UK quantitative WB-DWI technical workgroup: Consensus meeting recommendations on optimisation, quality control, processing and analysis of quantitative whole-body diffusion-weighted imaging for cancer. *Br. J. Radiol.* **2018**, *91*, 20170577. [[CrossRef](#)]
19. Winfield, J.M.; Tunariu, N.; Rata, M.; Miyazaki, K.; Jerome, N.P.; Germuska, M.; Blackledge, M.D.; Collins, D.J.; de Bono, J.S.; Yap, T.A.; et al. Extracranial Soft-Tissue Tumors: Repeatability of Apparent Diffusion Coefficient Estimates from Diffusion-weighted MR Imaging. *Radiology* **2017**, *284*, 88–99. [[CrossRef](#)] [[PubMed](#)]
20. Blackledge, M.D.; Tunariu, N.; Orton, M.R.; Padhani, A.R.; Collins, D.J.; Leach, M.O.; Koh, D.-M. Inter- and Intra-Observer Repeatability of Quantitative Whole-Body, Diffusion-Weighted Imaging (WBDWI) in Metastatic Bone Disease. *PLoS ONE* **2016**, *11*, e0153840. [[CrossRef](#)] [[PubMed](#)]
21. Reischauer, C.; Froehlich, J.M.; Koh, D.M.; Graf, N.; Padevit, C.; John, H.; Binkert, C.A.; Boesiger, P.; Gutzeit, A. Bone metastases from prostate cancer: Assessing treatment response by using diffusion-weighted imaging and functional diffusion maps - Initial observations. *Radiology* **2010**, *257*, 523–531. [[CrossRef](#)]
22. Blackledge, M.D.; Collins, D.J.; Tunariu, N.; Orton, M.R.; Padhani, A.R.; Leach, M.O.; Koh, D.-M. Assessment of Treatment Response by Total Tumor Volume and Global Apparent Diffusion Coefficient Using Diffusion-Weighted MRI in Patients with Metastatic Bone Disease: A Feasibility Study. *PLoS ONE* **2014**, *9*, e91779. [[CrossRef](#)]
23. Perez-Lopez, R.; Lorente, D.; Blackledge, M.D.; Collins, D.J.; Mateo, J.; Bianchini, D.; Omlin, A.; Zivi, A.; Leach, M.O.; De Bono, J.S.; et al. Volume of bone metastasis assessed with whole-Body Diffusion-weighted imaging is associated with overall survival in metastatic castration-resistant prostate cancer. *Radiology* **2016**, *280*, 151–160. [[CrossRef](#)]
24. Perez-Lopez, R.; Mateo, J.; Mossop, H.; Blackledge, M.D.; Collins, D.J.; Rata, M.; Morgan, V.A.; Macdonald, A.; Sandhu, S.; Lorente, D.; et al. Diffusion-weighted imaging as a treatment response biomarker for evaluating bone metastases in prostate cancer: A pilot study. *Radiology* **2017**, *283*, 168–177. [[CrossRef](#)] [[PubMed](#)]
25. Blackledge, M.D.; Leach, M.O.; Collins, D.J.; Koh, D.M. Computed diffusion-weighted MR imaging may improve tumor detection. *Radiology* **2011**, *261*, 573–581. [[CrossRef](#)] [[PubMed](#)]
26. Zhang, H.; Xue, H.; Alto, S.; Hui, L.; Kannengiesser, S.; Berthold, K.; Jin, Z. Integrated Shimming Improves Lesion Detection in Whole-Body Diffusion-Weighted Examinations of Patients With Plasma Disorder at 3 T. *Investig. Radiol.* **2016**, *51*, 297–305. [[CrossRef](#)]
27. Padhani, A.R.; Van Ree, K.; Collins, D.J.; D'Sa, S.; Makris, A. Assessing the relation between bone marrow signal intensity and apparent diffusion coefficient in diffusion-weighted MRI. *Am. J. Roentgenol.* **2013**, *200*, 163–170. [[CrossRef](#)] [[PubMed](#)]

28. Messiou, C.; Collins, D.J.; Morgan, V.A.; Desouza, N.M. Optimising diffusion weighted MRI for imaging metastatic and myeloma bone disease and assessing reproducibility. *Eur. Radiol.* **2011**, *21*, 1713–1718. [[CrossRef](#)]
29. Lavdas, I.; Rockall, A.G.; Castelli, F.; Sandhu, R.S.; Papadaki, A.; Honeyfield, L.; Waldman, A.D.; Aboagye, E.O. Apparent Diffusion Coefficient of Normal Abdominal Organs and Bone Marrow from Whole-Body DWI at 1.5 T: The Effect of Sex and Age. *Am. J. Roentgenol.* **2015**, *205*, 242–250. [[CrossRef](#)]
30. Messiou, C.; Collins, D.J.; Giles, S.; de Bono, J.S.; Bianchini, D.; de Souza, N.M. Assessing response in bone metastases in prostate cancer with diffusion weighted MRI. *Eur. Radiol.* **2011**, *21*, 2169–2177. [[CrossRef](#)]
31. Zou, K.H.; Warfield, S.K.; Bharatha, A.; Tempany, C.M.C.; Kaus, M.R.; Haker, S.J.; Wells, W.M.; Jolesz, F.A.; Kikinis, R. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Acad. Radiol.* **2004**, *11*, 178–189. [[CrossRef](#)]
32. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307–310. [[CrossRef](#)]
33. Bland, J.M.; Altman, D.G. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135–160. [[CrossRef](#)]
34. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [[CrossRef](#)]
35. Messiou, C.; Hillengass, J.; Delorme, S.; Lecouvet, F.E.; Mouloupoulos, L.A.; Collins, D.J.; Blackledge, M.D.; Abildgaard, N.; Østergaard, B.; Schlemmer, H.-P.; 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. [[CrossRef](#)] [[PubMed](#)]
36. Cui, F.-Z.; Cui, J.-L.; Wang, S.-L.; Yu, H.; Sun, Y.-C.; Zhao, N.; Cui, S.-J. Signal characteristics of normal adult bone marrow in whole-body diffusion-weighted imaging. *Acta Radiol.* **2016**, *57*, 1230–1237. [[CrossRef](#)] [[PubMed](#)]
37. Chen, Y.-Y.; Wu, C.-L.; Shen, S.-H. High Signal in Bone Marrow on Diffusion-Weighted Imaging of Female Pelvis: Correlation With Anemia and Fibroid-Associated Symptoms. *J. Magn. Reson. Imaging* **2018**, *48*, 1024–1033. [[CrossRef](#)] [[PubMed](#)]
38. Dionísio, F.C.F.; Oliveira, L.S.; Hernandes, M.A.; Engel, E.E.; Rangayyan, R.M.; Azevedo-Marques, P.M.; Nogueira-Barbosa, M.H. Manual and semiautomatic segmentation of bone sarcomas on MRI have high similarity. *Braz. J. Med. Biol. Res.* **2020**, *53*, e8962. [[CrossRef](#)] [[PubMed](#)]
39. Chen, L.; Sun, P.; Hao, Q.; Yin, W.; Xu, B.; Ma, C.; Stemmer, A.; Fu, C.; Wang, M.; Lu, J. Diffusion-weighted MRI in the evaluation of the thyroid nodule: Comparison between integrated-shimming EPI and conventional 3D-shimming EPI techniques. *Oncotarget* **2018**, *9*, 26209–26216. [[CrossRef](#)] [[PubMed](#)]