Magnetic Resonance Materials in Physics, Biology and Medicine (2025) 38:593–609
https://doi.org/10.1007/s10334-025-01256-0

RESEARCH ARTICLE

Multi-center and multi-vendor evaluation study across 1.5 T and 3 T scanners (part 1): apparent diffusion coefficient standardization in a diffusion MRI phantom

Siria Pasini[1] · Steffen Ringgaard[2] · Tau Vendelboe[2] · Leyre Garcia-Ruiz[3] · Anika Strittmatter[4,5] · Giulia Villa[1] · Anish Raj[4,5] · Rebeca Echeverria-Chasco[3] · Michela Bozzetto[1] · Paolo Brambilla[6] · Malene Aastrup[2] · Esben S. S. Hansen[2] · Luisa Pierotti[7] · Matteo Renzulli[8] · Susan T. Francis[9] · Frank G. Zoellner[4,5] · Christoffer Laustsen[2] · Maria A. Fernandez-Seara[3] · Anna Caroli[1]

Received: 24 October 2024 / Revised: 21 March 2025 / Accepted: 15 April 2025 / Published online: 9 May 2025

Abstract

Objective To validate multi-site and multi-vendor ADC measurements using the QIBA/NIST diffusion MRI phantom at room temperature.

Materials and methods ADC measurements were performed on 12 scanners (evenly split between 1.5 and 3 T) from three vendors at five sites and compared with reference values at room temperature. We adopted Pearson's correlation (r) and accuracy error for comparison with reference values; within scanner coefficient of variation (CV_{intra}%) for intra-session repeatability and inter-scanner for agreement (CV_{inter}%); Bland–Altman plots and precision error for short-term reproducibility; generalized linear mixed models and post-hoc tests (α=0.05) to compare accuracy, repeatability and precision across field strengths, vendors, and scanners.

Results Temperature adjusted ADCs were well correlated with NIST reference values ($r \geq 0.997$ for 1.5 T, $r \geq 0.996$ for 3 T). Median accuracy error was lower than 5% for all scanners. In the renal physiologic range ($ADC > 0.83 \times 10^{-3}$ mm^2/s), accuracy error was < 10% and $CV_{intra} < 2\%$. Across all scanners, good short-term reproducibility with limits of agreement < 10% and excellent agreement (median $CV_{inter} < 2\%$) were found.

Discussion Despite using abdominal receive coils and room temperature measurements, all quantitative parameters were within literature findings. High accuracy, repeatability and precision within the renal physiologic range support the feasibility of scanner evaluation using QIBA standardization process for diffusion measurements in renal studies.

Keywords Phantom · MRI · Apparent diffusion coefficient · Multi-site · Standardization

✉ Anna Caroli
anna.caroli@marionegri.it

[1] Bioengineering Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Camozzi 3, 24020 Ranica, BG, Italy

[2] The MR Research Centre, Aarhus University, Aarhus, Denmark

[3] Department of Radiology, Clínica Universidad de Navarra, Pamplona, Spain

[4] Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[5] Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[6] Radiology Unit, ASST Papa Giovanni XXIII, Bergamo, Italy

[7] Department of Medical Physics, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

[8] Department of Radiology, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

[9] Sir Peter Mansfield Imaging Centre, University of Nottingham, Nottingham, UK

Introduction

Diffusion-weighted imaging (DWI) is a non-invasive technique in magnetic resonance imaging (MRI) to evaluate the local displacement of water molecules in tissues [1]. Quantitatively, this is expressed in terms of apparent diffusion coefficient (ADC), a functional parameter derived from fitting a mono-exponential model to signal intensity acquired at varying diffusion weighting. The potential of ADC as an effective imaging biomarker has been recognized in many clinical applications [2, 3] mainly for its capability of non-invasively probing and characterizing tissue microstructures. In particular, for renal imaging applications, DWI shows potential to detect and monitor interstitial fibrosis in chronic kidney disease, as well as inflammation in acute kidney diseases [4].

However, there are limited studies on normal ranges for ADC in the kidney. Indeed, diffusion metrics have been shown to vary significantly depending on the type and field strength of scanner, sequence and data analysis method employed [5], hindering reliable comparisons in multi-center MRI clinical studies and the ultimate translation of quantitative DWI biomarkers to clinical practice. To standardize ADC, the Quantitative Imaging Biomarker Alliance (QIBA) [6] alongside the National Institute of Standards and Technology (NIST) has developed a commercially available ADC phantom which can be used with standardized vendor-specific protocols [7]. Scanner performances can therefore be assessed consistently [8], in terms of accuracy, repeatability and reproducibility [9], both with the QIBA technical standards and across platforms.

While previous phantom studies [5, 8, 10–12] have performed the QIBA conformance process [13] with the QIBA/NIST phantom at 0 °C with the recommended head coil, there have been few studies at room temperature [14, 15] or using alternative RF coils, particularly at a multi-site level. For renal studies, this implies using surface body/torso coils and keeping the reference temperature at room temperature close to 20 °C to assess ADC over a higher range compatible with the relevant physiologic range for kidneys [16–18]. In healthy kidneys, published ADC values vary from 2.0–4.1 $\times 10^{-3}$ mm^2/s in the cortex, 1.9–5.1 $\times 10^{-3}$ mm^2/s in the medulla and 1.6–3.7 $\times 10^{-3}$ mm^2/s in the whole parenchyma [16, 18]. As highlighted by Zhang et al. [16] this broad range mainly arises due to differences in imaging parameters used and unphysical diffusion values (ADC > 3 $\times 10^{-3}$ mm^2/s) likely resulted from the unsuppressed IVIM contribution at low b-values (< 200 s/mm^2) [19]. Suo et al. [18], using b-values outside the low regime (0–600 s/mm^2), reported mean physiologic ADC values of 2.31 $\times 10^{-3}$ mm^2/s in the cortex, 2.15 $\times 10^{-3}$ mm^2/s in the medulla and 2.23 $\times 10^{-3}$ mm^2/s in the whole parenchyma across 137 healthy patients. Similarly, Xu et al. [20], in a study with 72 healthy volunteers using b-values of 0–500 s/mm^2, estimated a mean ADC value in the kidneys 2.45–2.52 $\times 10^{-3}$ mm^2/s. For CKD patients, ADC values are reported as significantly lower than healthy ones ranging from 0.88–2.2 $\times 10^{-3}$ mm^2/s depending on the stage of the disease [17, 20, 21].

The aim of this study, performed in the context of a multi-center project focused on renal MRI standardization to improve personalized management of patients with chronic kidney disease (RESPECT, https://respectmri.com), was to enhance ADC reproducibility by adhering to QIBA conformance procedures for ADC measurements across multiple sites and scanner vendors, using ambient temperature conditions and scanner set-ups typical of abdominal studies. Field strength dependency was also explored by including both 1.5 T and 3 T scanners from each site. Furthermore, characterization of three orthogonal imaging planes was performed for all scanners. All results were compared with NIST-provided reference values.

Materials and methods

Diffusion phantom

ADC values were measured on the QIBA Diffusion Phantom (Fig. 1a) (Model 128, CaliberMRI, Boulder, CO, USA). For the study, the same phantom was scanned across all sites participating in diffusion data collection. The phantom consists of a spherical shell with an outer diameter of 19.4 cm containing a single central vial plate that holds 13 cylindrical vials. The vials are filled with 30 mL of aqueous polyvinylpyrrolidone (PVP) solution at increasing concentrations [22] (0, 10, 20, 30, 40, 50%). The location of the vials and their corresponding PVP concentration is shown in Fig. 1b. Reference values were provided by NIST Boulder at temperatures of 16, 18, 20, 22, 24, and 26 °C [23].

In addition, the phantom contains 10 liquid crystal (LC) MR-readable thermometer vials covering the temperature range from 15 to 24 °C as shown in Fig. 1d.

Image acquisition

The diffusion phantom was scanned by 6 different sites at both 3 T and 1.5 T (Table 1), on scanners from the three dominant MRI vendors ($n = 5$ from Siemens Healthineers AG (Forchheim, Germany), 4 from GE Healthcare (Waukesha, WI), and 3 from Philips Healthcare (Best, the Netherlands)). Each acquisition was repeated on a different day for the assessment of intra-site short-term

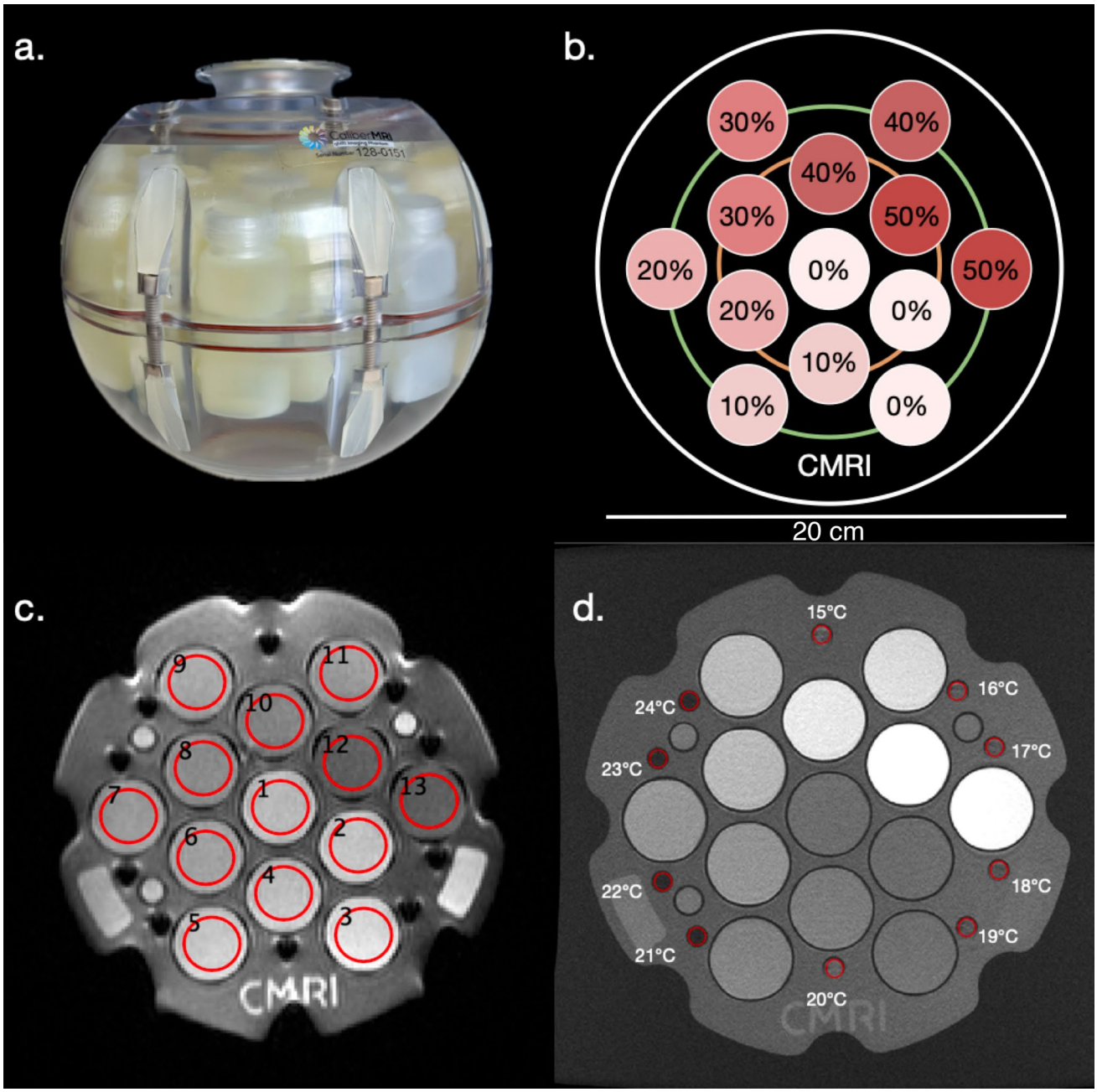


Fig. 1 **a** Side view of the CaliberMRI Diffusion phantom. **b** Diagram of the central plate coronal view showing the vials location and PVP concentration. The vials have PVP concentrations ranging from 0 to 50% and they are positioned clockwise with increasing concentration along two concentric rings (inner ring represented in orange and outer ring in green). The central vial contains high-purity water (0% PVP concentration). **c**, **d** Coronal view of the central plate showing ROIs size and positioning both for ADC (**c**) and temperature (**d**) estimation

reproducibility. For all scanners, DWI trace images were obtained using three orthogonal-direction single-shot echo planar imaging (SS-EPI) DWI protocols with 4 b-values (0, 500, 900, 2000 s/mm^2) as suggested by the phantom manufacturer (CaliberMRI) [22] (Table 2). For each scanner, imaging was performed with the available abdomen and spine receive coils as reported in Table 2.

To minimize the effect of gradient non-linearities [11], during each acquisition, the phantom was positioned with the central vial at the scanner isocenter. Because of the phantom spherical shape, movements were minimized by using cushions. DWI images were acquired along three diffusion directions and with slices oriented along each of the three orthogonal imaging planes (coronal, axial and sagittal) after physically rotating the phantom as indicated in

Table 1 MRI scanner details

Scanner ID	Site	Field strength (T)	Vendor	Vendor ID	Scanner type	Phantom temperature (°C)	
						Scan 1	Scan 2
1	IRFMN	1.5	GE	A	Optima 450w	20.5 ± 0.2	20.6 ± 0.4
2	UNAV	1.5	Siemens	B	Aera	24.0*	22.7 ± 0.6
3	UHEI	1.5	Siemens	B	Aera	20.0 ± 0.4	22.0 ± 0.5
4	AUH	1.5	GE	A	Optima 450W	20.7 ± 0.5	20.5 ± 0.2
5	AUH	1.5	Philips	C	Achieva Stream	20.5 ± 0.6	20.5 ± 0.5
6	UNIBO	1.5	Philips	C	Ingenia	20.1 ± 0.5	20.3 ± 0.3
7	IRFMN	3	GE	A	Discovery 750W	20.6 ± 0.5	20.2 ± 0.5
8	UNAV	3	Siemens	B	Skyra	22.0 ± 0.4	23.0 ± 0.2
9	UHEI	3	Siemens	B	Skyra	21.8 ± 0.5	22.0 ± 0.4
10	AUH	3	GE	A	Discovery 750	20.0 ± 0.2	19.9 ± 0.4
11	AUH	3	Siemens	B	Skyra	20.6 ± 0.6	20.0 ± 0.4
12	UNIBO	3	Philips	C	Ingenia	20.6 ± 0.5	20.8 ± 0.5

IRFMN Istituto di Ricerche Farmacologiche Mario Negri, *UNAV* Clínica Universidad de Navarra, *UHEI* Heidelberg University, *AUH* Aarhus University, *UNIBO* Azienda ospedaliero-universitaria di Bologna

*Maximum temperature that could be estimated using the LC MR-readable thermometer (see Supplementary Fig. S4)

Table 2 Sequence details

Parameters	GE	Siemens	Philips
Field strength (T)	1.5 T and 3 T	1.5 T and 3 T	1.5 T and 3 T
Acquisition sequence	SS-EPI	SS-EPI	SS-EPI
Receive coil type	32–48 channel body coil	32 channel body coil	Body MULTI-COIL
b-values [s/mm^2]	0, 500, 900, 2000	0, 500, 900, 2000	0, 500, 900, 2000
Diffusion directions	3	3	3
Slice/ Gap thickness [mm]	4/1	4/1	4/1
Field of view [mm]	220 × 220	216 × 220	220 × 220
Acquisition matrix	128 × 128	184 × 186	128 × 128 (3 T) 128 × 126 (1.5 T)
Reconstruction matrix	256 × 256	184 × 186	256 × 256
Plane orientation	1st Coronal; 2nd Axial; 3rd Sagittal	1st Coronal; 2nd Axial; 3rd Sagittal	1st Coronal; 2nd Axial; 3rd Sagittal
Number of averages	1	1	1
In-plane parallel imaging acceleration factor	2	2	2
TR/TE [ms]	TR = 10,000 TE = Minimum (90 ms, 110 ms for Scanner 4)	TR = 10,000 TE = Shortest (110 ms)	TR = 10,000 TE = Minimum (110 ms, 81 ms for Scanner 6)
Voxel size [mm^3]	0.8594 × 0.8594 × 5	1.1828 × 1.1828 × 5	0.8594 × 0.8594 × 5

SS-EPI Single shot echo planar imaging, *TR* repetition time, *TE* echo time

the CaliberMRI phantom manual (Supplementary Fig. S1). For each imaging plane, four sequential DWI images were acquired without changing acquisition conditions and calibration.

The phantom thermalized in the scanner room for at least 24 h before each acquisition. Temperature measurements for each acquisition were obtained using an isotropic spoiled gradient echo sequence on the MR-readable thermometer embedded in the phantom [24]. Temperature

determination was performed before and after each data set so that any change in temperature could be monitored.

Image processing

Image analysis was conducted uniformly across all sites by a single operator (S.P.), ensuring consistency in the methodology employed [5, 25]. Image processing was performed with PhantomViewer software [26]. For all acquisitions, only the central slice across the vials was used in the analyses. Circular regions of interest (ROIs) with an average diameter of 20 mm were manually placed on each vial, edges were avoided during placement (Fig. 1c). Given the resolution of each acquisition, each ROI consisted of more than 200 pixels. Mean ROI signal intensity was obtained for each vial. ROI size and position

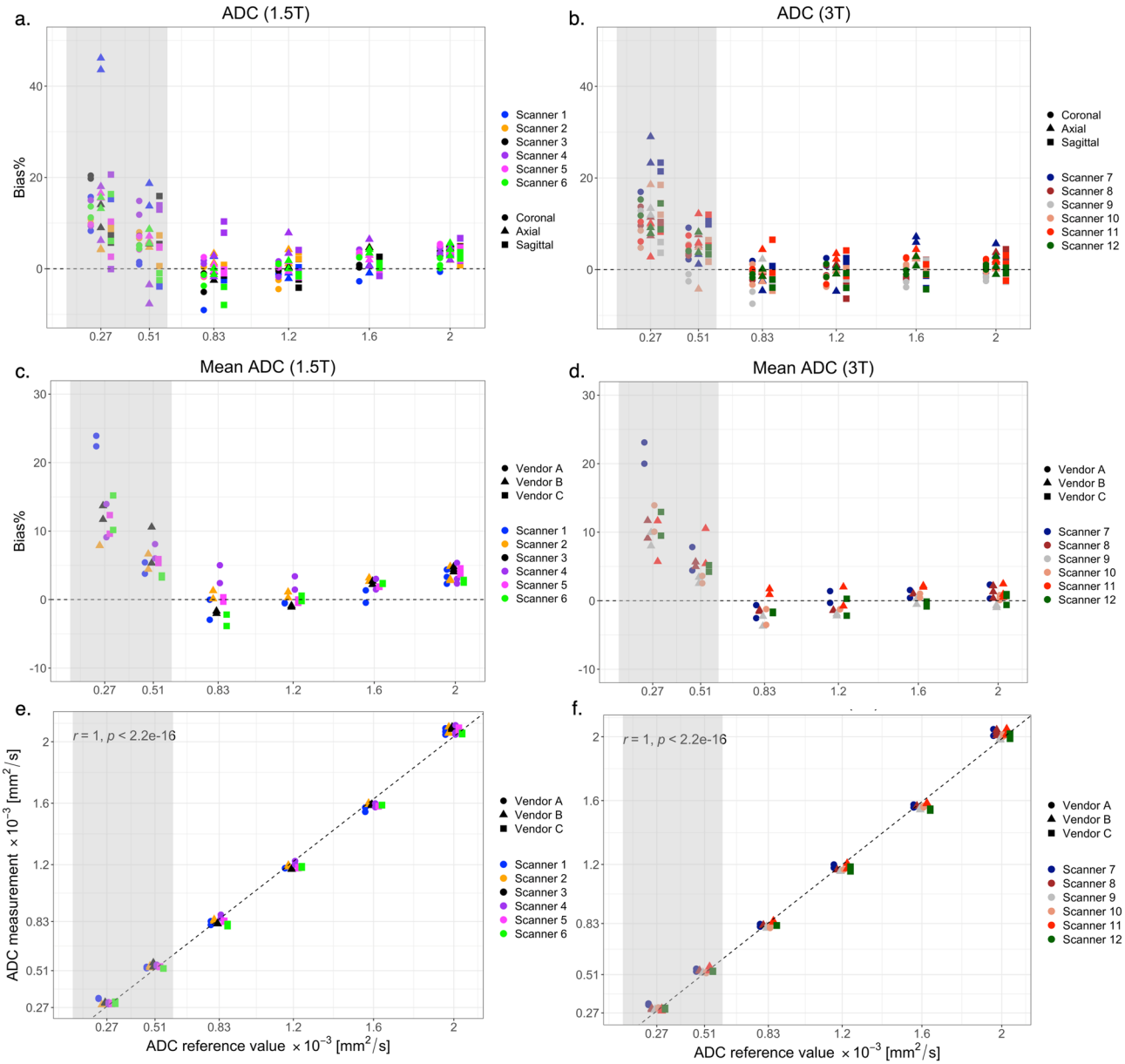


Fig. 2 Bias% of temperature corrected ADC values plotted with respect to ADC reference along each imaging plane for both 1.5 T scanners (**a**) and 3 T scanners (**b**). Average bias% across the three imaging planes for 1.5 T scanners (**c**) and 3 T scanners (**d**). Horizontal dashed lines represent 0% bias. Correlation plots of average temperature corrected ADC values and ADC reference for 1.5 T scanners (**e**) and 3 T scanners (**f**). Correlation lines obtained from linear regression are represented as black dashed lines. Pearson's correlation coefficients with their corresponding p value are reported in the upper left corner (**e**–**f**). The area shaded in gray corresponds to the values outside of the renal ADC physiologic range

were adjusted only in the presence of visible artifacts to avoid an incorrect estimation of the average signal. For Philips scanners, signal intensity values were properly rescaled [27] before proceeding with the calculations. Curve fitting was performed in Python through non-linear least-square minimization with parameters bounded within their physically acceptable constraints. For each vial, ADC values were calculated with a mono-exponential fit as given by $S(b) = S_0 e^{-b \bullet ADC} + n$, with initial values $S_0 = \max(S)$ and constraints $S_0 \geq 0$ and $0 \leq n \leq \min(S)$. For all fits, the starting point for baseline noise was estimated from the background signal as described by Fang et al. [14]. Four circular background ROIs (3 cm of diameter) were positioned at each corner outside of the phantom. ADC values were separately determined for each repetition and along each imaging plane; an average value was then calculated both characterizing each imaging plane and overall to obtain a spatially independent mean value $\overline{ADC}$ for each vial.

Signal to noise ratio and signal homogeneity

For all scanners and imaging planes, the signal to noise ratio (SNR) was calculated with the difference method [6, 28] (SNR_{diff}). This method also provided an initial estimate of the baseline noise for the mono-exponential fit by determining the standard deviation of the difference image signal in the center [14] of the phantom. SNR_{diff} agreement between imaging planes was tested with BA statistics. In addition, when sequential repetitions of trace images were available, SNR was calculated as the ratio between the temporal mean and temporal standard deviation images (SNR_{mult}) [6, 28]. For those scanners which provided only two sequential acquisitions, SNR_{mult} was estimated from two symmetrically positioned slices taken before and after the central slice. Both these methods are independent of the noise spatial distribution statistics thus can account for coil geometry, reconstruction methods and parallel imaging [28]. SNR was calculated separately for each ROI and *b*-value.

Signal homogeneity along each imaging plane and for each scanner was characterized by measuring the ROI signal ratio% of all vials with the same PVP concentrations either between the inner and outer ring (Fig. 1b) or, for the 0% vial, between the center and the inner ring (Supplementary Fig. S2). Furthermore, signal homogeneity within each ROI was estimated for each b-value in terms of percentage image uniformity (PIU) [29] (Fig. S2). This part of the analysis was performed with Python 3.11.7.

Pure water vials (PVP = 0%), positioned in the isocenter, the inner circle (~ 4 cm from the isocenter) and the outer circle (~ 8 cm from the isocenter), were also used to investigate the effect of radial distance from the isocenter on ADC measurements. Spatially dependent bias was measured for each scanner as the relative difference % between the inner/outer ADC and the isocenter ADC value [30] (Supplementary Fig. S3).

Temperature correction

Temperature estimation for each acquisition was performed with PhantomViewer software by placing ordered circular ROIs (5.5 mm of diameter) over the LC MR-readable thermometer vials (15, 16, 17, 18, 19, 20, 21, 22, 23, and 24 °C) as reported in Fig. 1d. Since ADC values exhibit temperature dependence [24] at both field strengths, to account for temperature variability, a valid comparison across scanners and

Table 3 Intra-session median repeatability and accuracy error%, in brackets is the range given by the lower and upper quartile

Scanner	Repeatability CV_{intra}%			Accuracy Error %			
	Coronal	Axial	Sagittal	Coronal	Axial	Sagittal	$\overline{ADC}$
1	1.06 (0.60–1.31)	0.86 (0.60–1.27)	1.07 (0.47–1.89)	2.53 (1.34–3.44)	2.78 (1.87–13.74)	2.51 (2.08–4.84)	2.95 (0.56–4.38)
2	0.30 (0.25–0.75)	0.28 (0.22–0.61)	0.56 (0.35–0.67)	4.53 (3.03–6.67)	4.17 (2.79–4.67)	1.83 (0.85–4.96)	2.96 (1.30–4.82)
3	0.44 (0.29–0.54)	0.26 (0.18–0.45)	0.37 (0.18–0.61)	3.88 (0.84–5.13)	4.61 (2.52–5.42)	4.13 (2.37–5.42)	4.05 (2.03–5.33)
4	0.81 (0.39–1.04)	0.86 (0.46–1.69)	0.48 (0.18–1.03)	3.37 (1.83–9.37)	3.51 (2.86–6.44)	4.11 (1.58–10.35)	3.40 (2.41–6-01)
5	0.42 (0.32–0.68)	0.50 (0.23–0.56)	0.85 (0.70–1.01)	4.20 (2.55–5.42)	3.62 (1.58–5.80)	1.89 (1.00–4.64)	3.67 (0.48–5.33)
6	0.32 (0.23–0.49)	0.40 (0.28–0.42)	0.19 (0.17–0.26)	2.46 (1.14–4.39)	4.19 (1.77–5.61)	2.50 (1.27–3.97)	2.79 (2.29–3.52)
7	0.19 (0.11–0.29)	0.37 (0.28–0.60)	0.15 (0.13–0.21)	1.89 (0.83–2.51)	4.59 (2.41–5.96)	2.51 (1.35–9.87)	2.32 (0.64–4.43)
8	0.20 (0.14–0.25)	0.29 (0.17–0.32)	0.18 (0.16–0.47)	1.67 (1.00–3.06)	2.41 (0.64–5.27)	3.35 (2.12–6.32)	1.42 (1.31–4.99)
9	0.25 (0.15–0.31)	0.31 (0.23–0.45)	0.15 (0.06–0.39)	2.68 (1.80–4.81)	2.21 (0.90–5.94)	2.64 (2.02–3.97)	2.21 (0.74–3.43)
10	0.29 (0.27–0.38)	0.24 (0.17–0.45)	0.69 (0.52–1.26)	1.76 (1.11–4.34)	1.68 (0.84–4.24)	1.68 (0.63–4.61)	1.23 (0.82–3.52)
11	0.15 (0.13–0.28)	0.40 (0.21–0.65)	0.26 (0.09–0.38)	2.46 (1.65–5.22)	2.94 (2.10–4.40)	3.14 (1.22–6.51)	2.01 (0.93–5.42)
12	0.19 (0.15–0.27)	0.29 (0.17–0.47)	0.18 (0.17–0.34)	1.26 (1.00–4.04)	1.50 (0.79–3.75)	3.36 (0.80–4.29)	1.61 (0.71–4.23)

CV_{intra} intra-session coefficient of variation

with reference values was performed after adjusting each measurement to a common reference temperature (20 °C) [31]. Details on temperature estimation and temperature adjustment are reported in Supplementary Material (Figs. S4, S5).

Statistical analysis

Statistical analysis was performed using R Studio (version 4.2.1) [32], both in the full range covered by the phantom vials ($0.27–2.0 \times 10^{-3}$ mm^2/s) and in the physiologically relevant range for kidneys (ADC $> 0.83 \times 10^{-3}$ mm^2/s corresponding to PVP concentration < 40%).

The agreement between the temperature adjusted ADC measurements and the provided NIST reference at 20 °C was evaluated by calculating their relative difference (bias%). Accuracy error [25] was calculated as the absolute value of bias.

$$\text{Accuracy Error\%} = 100 \bullet \frac{\left|\text{ADC}_{\text{adjusted}} - \text{ADC}_{\text{reference}}\right|}{\text{ADC}_{\text{reference}}}.$$

Accuracy analysis was performed both on the intra-session average along each imaging plane and on the overall position-independent average $\overline{ADC}$. For each scanner, accuracy error is reported as its median value along with the lower and upper quartiles. The relation between accuracy error and the inverse of SNR_{diff} was tested by Pearsons's correlation coefficient (r). The same test was also adopted to determine the concordance of ADC measurements between imaging planes (coronal-axial, coronal-sagittal) and to assess the agreement of $\overline{ADC}$ with respect to the reference values.

Intra-session repeatability for each imaging plane was studied in terms of coefficient of variation CV_{intra}% [6].

$$\text{CV}_{\text{intra}}\% = 100 \bullet \frac{\sigma_w}{\mu_w}.$$

Higher values of CV_{intra}% represent lower intra-session repeatability. The correlation between the coefficient of variation across multiple repetitions and SNR_{mult} was tested by Pearsons's correlation coefficient (r). For each scanner CV_{intra} is reported as median values along with lower and upper quartiles.

For each scanner, short-term reproducibility was visually assessed by BA plots and quantitatively measured by Bland–Altman statistics and precision error [25]. BA results are reported in terms of average bias% between the first and the second scan and 95% limits of agreement. Precision error was computed as the percentage ratio between the first and second scan difference and their mean value.

$$\text{Precision error\%} = 100 \bullet \frac{\left|\text{ADC}_{\text{scan1}} - \text{ADC}_{\text{scan2}}\right|}{\text{Mean}\left(\text{ADC}_{\text{scan1}}, \text{ADC}_{\text{scan2}}\right)}.$$

Precision analysis was performed both on the average along each imaging plane and on $\overline{ADC}$ values.

A generalized linear mixed model [33] (GLMM) was used to compare accuracy error, repeatability and precision error across field strengths and vendors for each imaging plane direction by taking into account repeated measures [25]. The same analysis was repeated for accuracy and precision error calculated from $\overline{ADC}$ values. For each data field strength and vendor were introduced as fixed effects of the model and were analyzed both as independent predictors and in combination as two-way interaction. Scanner ID and vial ID were modeled as random effects. A GLMM was also used to compare accuracy error, repeatability and precision error between imaging plane orientations for each scanner. The best model was chosen by comparing model performance metrics using the R function "anova" from "stats" package [32]. Results are reported as least-square means and standard errors [25]. Statistical significance of predictors was expressed in terms of type 3 p values[25]. Post-hoc comparisons were performed for all predictors in the model with Bonferroni adjusted pairwise comparison ("emmeans" package). For all reported p values statistical significance was set at $p < 0.05$.

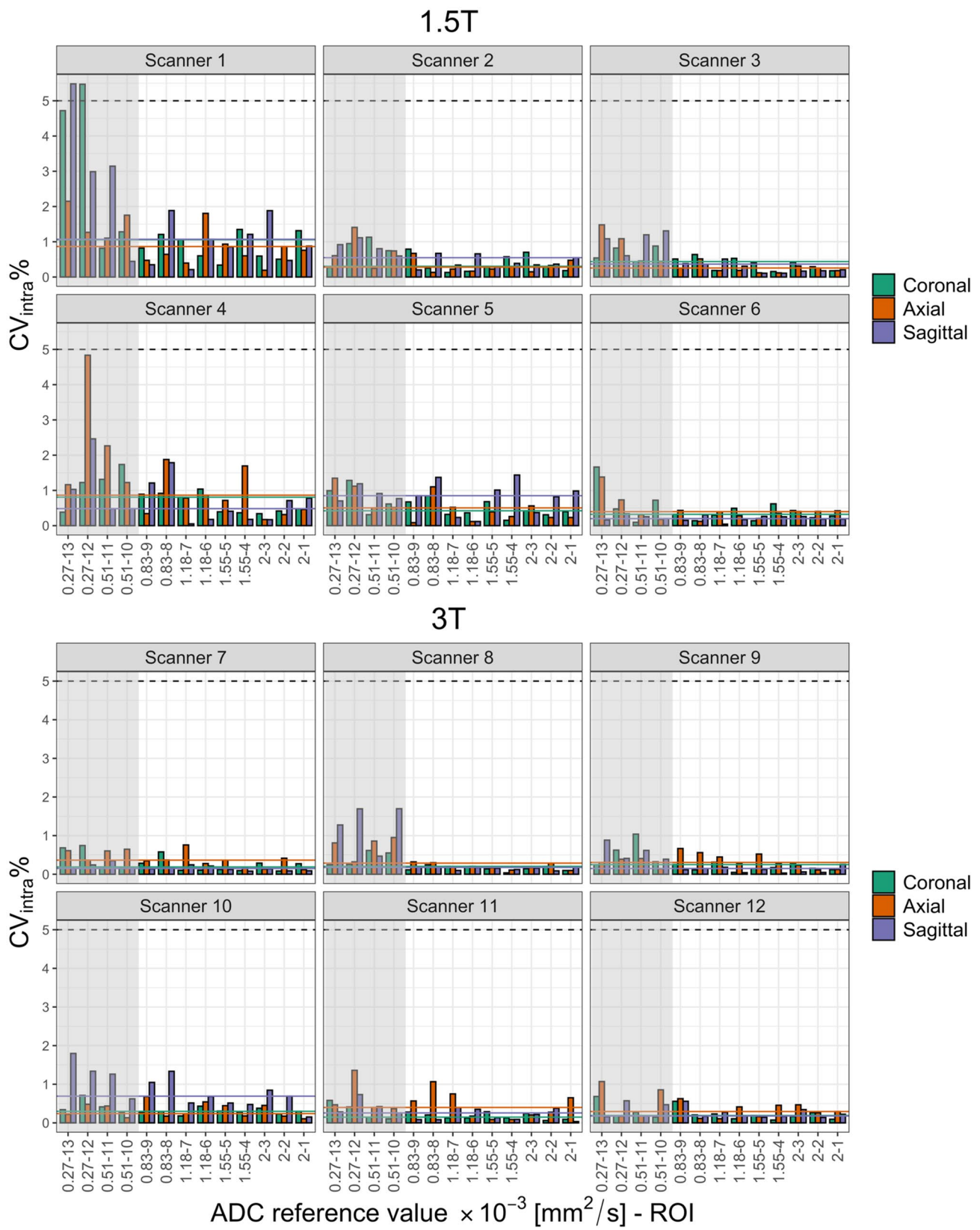
Inter-scanner agreement was measured by the coefficient of variation (CV_{inter}%), calculated as the ratio between the standard deviation and the average of the first session measurements across all scanners. CV_{inter}% was separately calculated for each vial both along each imaging plane and on the overall average $\overline{ADC}$ pooling data by field strength. Comparisons between groups were computed with paired sample Wilcoxon signed-rank tests.

For both field strengths and for each phantom vial, intra-site and inter-site contributions to the overall variance were analyzed by computing the intra-class and inter-class correlation coefficient ($\text{ICC}_{\text{intra}}$ and $\text{ICC}_{\text{inter}}$) [34].

$$\text{ICC}_{\text{intra}} = \frac{\sigma^2{}_{\text{intra}}}{\sigma^2{}_{\text{intra}} + \sigma^2{}_{\text{inter}}},$$

$$\text{ICC}_{\text{inter}} = \frac{\sigma^2{}_{\text{inter}}}{\sigma^2{}_{\text{intra}} + \sigma^2{}_{\text{inter}}}.$$

Intra-site variance $\sigma^2{}_{\text{intra}}$ was calculated as the average variance across sites, while inter-site variance $\sigma^2{}_{\text{inter}}$ was determined as the variance across averages for each site.


1.5T
Scanner 1
Scanner 2
Scanner 3
Scanner 4
Scanner 5
Scanner 6
3T
Scanner 7
Scanner 8
Scanner 9
Scanner 10
Scanner 11
Scanner 12
CV intra %
0 1 2 3 4 5
0.27-13 0.27-12 0.51-11 0.51-10 0.83-9 0.83-8 1.18-7 1.18-6 1.55-5 1.55-4 2-3 2-2 2-1
Coronal
Axial
Sagittal
ADC reference value × 10−3 [mm2/s] - ROI

◂**Fig. 3** Coefficient of variation (CV_{intra}%) of ADC values from sequential acquisitions calculated for each scanner at 1.5 T (**a**) and 3 T (**b**) for all three imaging planes. Horizontal solid lines represent CV_{intra}% median values along each imaging plane as depicted by the different colors while horizonal dashed lines in black represent the 5% threshold. The area shaded in gray corresponds to the values outside the renal ADC physiologic range

Results

Acquisitions were successfully performed for all scanners; however, a few deviations from the original protocol should be noted. One institution (Scanners 6 and 12) included a b-value of 1000 s/mm^2 instead of 900 s/mm^2 due to scanner technical requirements, and for Scanner 6 the acquisition matrix was lower than the suggested value in the manufacturer protocol (72 × 77 instead of 128 × 126). For Scanner 1, a different abdominal body coil was used between the first (24 channels body array coil) and the second scan (48 channels body array coil). One site (Scanners 4,5,10,11) included a lower number of repetitions (< 4) along axial and sagittal positioning. Finally, only for Scanner 12, SNR could not be determined due to signal scaling variations across repetitions.

Across all scanners, only Scanner 1 resulted in a SNR in the isocenter of < 30. Results of SNR calculations are reported in Supplementary Tables S1, S2. For this study, DWI images from all scanners had appropriate quality and significant artifacts were avoided adjusting ROI placement and size. Reduction in ROI size was performed ensuring a minimum acceptable dimension of 50 voxels per ROI.

Average temperature was determined for each scan and ranged from 19.9 to 24 °C (Table 1). For each scanner, temperature corrected ADC measurements across each imaging plane as well as an overall average are reported alongside the corresponding NIST reference ADC values, see Supplementary Table S3.

Results for spatially dependent bias are reported in Supplementary Fig. S3. Limits of agreement of relative difference% were within 5% for all inner vials except for sagittal positioning of 1.5 T scanner. For the outer vial, limits of agreement were within 8%. Among inner vials, 90% of ADC bias values were within the 3% expected negligible range [35], while for outer vials this value decreased to approximately 71%.

ADC measurements in comparison with NIST reference values

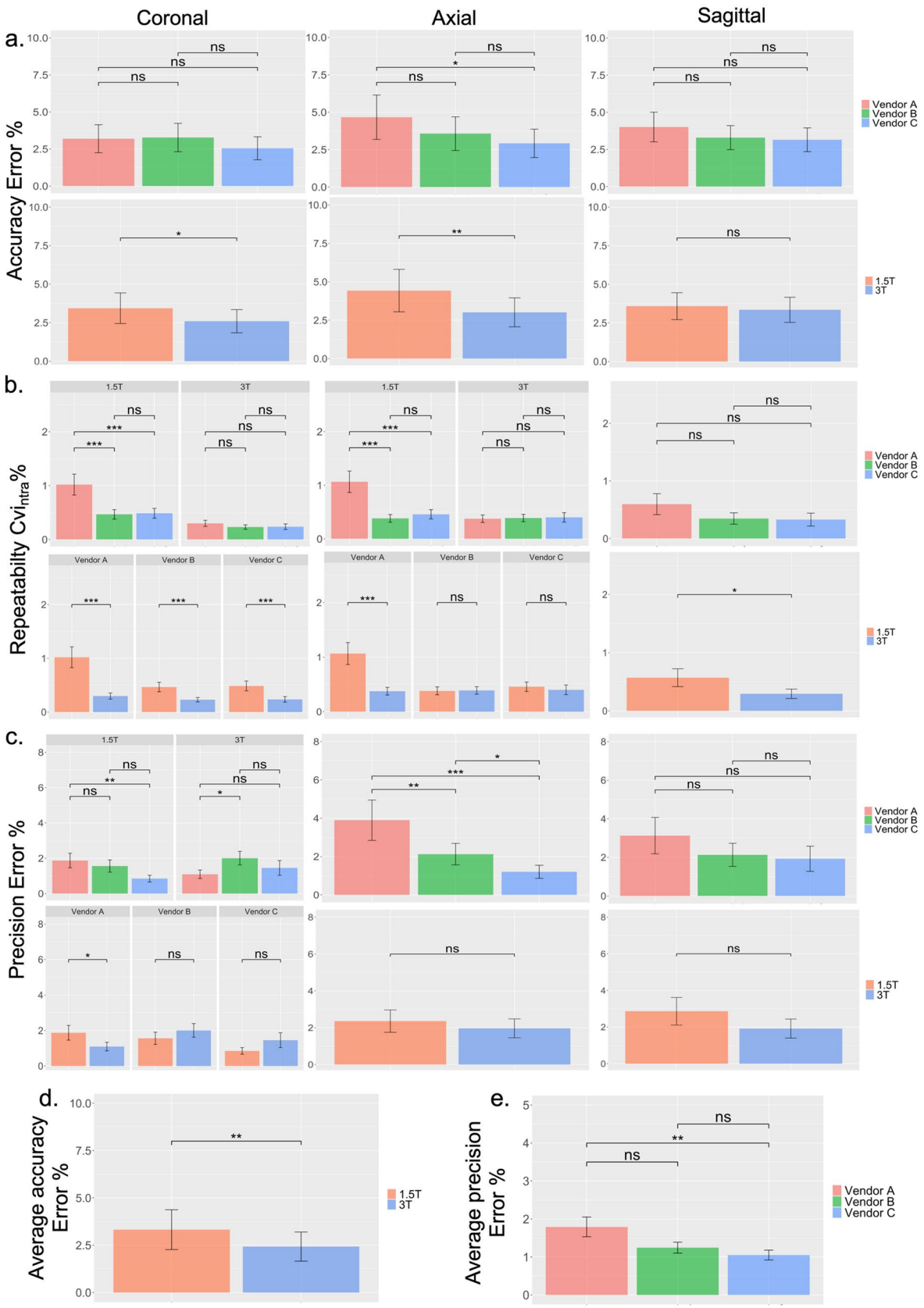
$\overline{ADC}$ Values had an excellent correlation with NIST reference values both for each scanner ($r > 0.999$) and pooling all scanners with the same field strength ($r = 1$) as shown in Fig. 2e, f. Similar results were obtained across each imaging plane. Excellent correlation was also found for ADC measurements of the same scanner across different imaging planes ($r \geq 0.997$ for 1.5 T, $r \geq 0.996$ for 3 T). For all scanners and imaging plane directions, bias% showed an asymmetrical distribution across the zero-bias line with an average positive bias for lower ADC values ($< 0.83 \times 10^{-3}$ mm^2/s). Among 1.5 T and 3 T scanners, the highest bias% was found for Scanner 1 and 7 along the axial direction ($bias_1 > 40\%$ and $bias_7 > 25\%$ at $0.2\,7 \times 10^{-3}$ mm^2/s) (Fig. 2a, b). This was reflected by $\overline{ADC}$ values, where the highest bias was found for Scanner 1 and 7 ($bias_{1,7} > 20\%$ at $0.2\,7 \times 10^{-3}$ mm^2/s) (Fig. 2c, d). In the renal physiologically relevant range, bias was within 10% both across each imaging plane and for the overall average.

ADC values in the low range (0.27–0.5 1×10^{-3} mm^2/s) were characterized by a larger data dispersion corresponding to a within imaging plane $CV_{inter} > 5\%$ for both values along axial (11.8–6.46%) and sagittal (5.34–6.61%) imaging plane direction on 1.5 T scanners and only for the lowest value (6.53% along axial and 5.51% along sagittal) for 3 T scanners. For both field strengths and across all imaging planes the highest ADC value (2.0×10^{-3} mm^2/s) corresponded to the lowest CV_{inter}% (< 2%).

ADC accuracy across scanners

Results of accuracy error for each imaging plane and for $\overline{ADC}$ are reported in Table 3. Median accuracy error among all scanners ranged from 1.26 to 4.61% for the position dependent results and from 1.23 to 4.05% for the averages. The highest median accuracy error was found among 1.5 T scanners: Scanner 2 along coronal positioning (4.53%), Scanner 3 along axial positioning (4.61%) and sagittal positioning (4.13%). Among position-independent results, 1.5 T scanners had higher median accuracy errors (range from 2.79 to 4.05%) compared to 3 T scanners (range from 1.23 to 2.32%). In the isocenter, 94% of ADC values were within the 5% range with the highest accuracy error found along sagittal imaging plane for both field strengths (6.71% on Scanner 4 and 3.14% on Scanner 11) (Supplementary Table S4). Accuracy error had a significant strong correlation ($r \geq 0.624$, $p \leq 2.16 \times 10^{-5}$) with the inverse of SNR_{diff} for Scanners 5, 6, 7, 8, 10 and 11 and a significant moderate correlation ($0.476 \leq r \leq 0.579$, $p \leq 2.17 \times 10^{-3}$) for Scanners 1, 2, 3, 4 and 9.

In a GLMM, no significant interaction was found between field strength and vendor ($p \geq 0.16$) thus both factors were introduced in the model as independent predictors for each imaging plane (Supplementary Table S5). Results showed that accuracy error was significantly higher for 1.5 T compared to 3 T scanners (Fig. 4a) across all imaging planes except sagittal positioning ($p = 0.55$). Among different vendors accuracy error was significantly higher for vendor A compared to vendor C ($p = 0.01$) only along axial positioning


Coronal
Axial
Sagittal
a.
Accuracy Error %
Vendor A
Vendor B
Vendor C
1.5T
3T
b.
Repeatabilty Cvi$_{Intra}$ %
c.
Precision Error %
d.
Average accuracy Error %
e.
Average precision Error %

◂**Fig. 4** Results of GLMM on accuracy error (**a**), repeatability (**b**) and precision error (**c**) across all imaging planes. Accuracy error and precision error are also reported for average values (**d**, **e**). All results are reported as least-square means and standard errors provided by GLMM. Horizontal lines represent the paired comparisons with their relative significance value (ns = $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

while no significant differences among vendors were found along coronal and sagittal direction. When comparing imaging plane positions (Supplementary Fig. S6), GLMM results showed that for most of the scanners no significant differences were found with some exceptions showing a significantly higher accuracy error for axial direction on scanner 1 ($p_{\text{cor-ax}} = 5.39 \times 10^{-3}$, $p_{\text{sag-ax}} = 0.03$) and 7 ($p_{\text{cor-ax}} = 0.01$) and sagittal direction on scanner 12 ($p_{\text{cor-sag}} = 2.05 \times 10^{-3}$, $p_{\text{sag-ax}} = 0.03$).

Accuracy error, obtained from $\overline{ADC}$ measurements, was also assessed with a GLMM (Fig. 4d). The best model showed no vendor–field interaction effect and a significant effect only from the field strength factor. Thus, when averaging across all imaging planes, no significant differences were found across vendors and a significantly higher accuracy error was present for 1.5 T scanners compared to 3 T scanners ($p = 5.38 \times 10^{-3}$).

In the renal relevant range median accuracy error, calculated for each scanner and along each direction, was reduced on average by approximately 30% compared to the full range. A similar reduction (27%) was obtained when considering position-independent accuracy error data. No differences were found in GLMM results across field strength and vendors when applied on the reduced range compared to the full range (Supplementary Fig. S7).

ADC repeatability

For each scanner the intra-session coefficient of variation across each imaging plane is reported in Fig. 3. Among 3 T scanners CV_{intra} was < 2% for all ROIs and imaging plane directions, among 1.5 T scanners this was true only for Scanners 2,3,5 and 6. Scanners 1 and 4 showed a CV_{intra}% that exceeded 2% only for ADC reference values in the lower range (0.27–0.51 × 10^{-3} mm^2/s). In the renal physiologic range, all scanners showed a CV_{intra}% lower than 2%. CV_{intra} had a significant strong correlation ($r \geq 0.637$, $p \leq 1.3 \times 10^{-5}$) with the inverse of SNR_{mult} for Scanners 1,2,3,4,7 and 10 and a significant moderate correlation ($0.460 \leq r \leq 0.568$, $p \leq 3.18 \times 10^{-3}$) for Scanners 5,6,8,9 and 11.

Among 3 T scanners, median CV_{intra}% (Table 3) ranged from 0.15 to 0.69%, with the highest value for Scanner 10 along the sagittal direction. Except for the highest value, all median CV_{intra}% were < 0.5%. Median CV_{intra} from 1.5 T scanners (Table 3) covered a higher range compared to 3 T scanners, from 0.19 to 1.07%. Along coronal and axial direction only Scanners 1 and 4 had a $CV_{\text{intra}} > 0.5\%$ while for sagittal positioning this was found for Scanners 1, 2 and 5. QIBA requirement on the isocenter coefficient of variation ($CV_{\text{intra}} < 0.5\%$), although originally defined for ice–water measurements, was verified for all 3 T temperature adjusted measurements along all imaging planes with one exception from axial direction of Scanner 11 (Supplementary Table S6). On 1.5 T scanners this was not the case for Scanner 1 along all directions and for sagittal positioning of Scanner 2, 4 and 5.

CV_{intra} for each imaging plane was tested with a GLMM in terms of two main factors (field strength and vendor) and their interaction (Fig. 4b). The interaction effect was found to be a significant predictor of repeatability CV_{intra}% along coronal and axial orientation ($p_{\text{cor}} = 0.03$, $p_{\text{ax}} = 1.96 \times 10^{-6}$, Table S5). 3T scanners, along coronal and axial positioning, showed no significant differences between vendors. For 1.5 T scanners vendor A had a significantly higher CV_{intra}% compared to other vendors along coronal ($p_{\text{A-B}} = 1.93 \times 10^{-6}$, $p_{\text{A-C}} = 7.42 \times 10^{-6}$) and axial ($p_{\text{A-B}} = 5.11 \times 10^{-10}$, $p_{\text{A-C}} = 4.27 \times 10^{-7}$) orientations. Across coronal and axial orientations 1.5 T scanners had a higher CV_{intra}% compared to 3 T scanners for vendor A ($p_{\text{cor}} = 4.29 \times 10^{-15}$, $p_{\text{ax}} = 7.49 \times 10^{-11}$). Only along coronal direction this difference was seen for vendor B and C as well ($p = 7.84 \times 10^{-7}$, $p = 1.52 \times 10^{-4}$). Across sagittal plane, no significant differences among vendors were found and 1.5 T scanners had a significantly higher CV_{intra}% compared to 3 T scanners ($p = 0.02$). Results of repeatability in the reduced range are reported Supplementary Fig. S7 and in Supplementary Fig. S8 for comparisons across scanners.

Average ADC inter-scanner agreement

Excellent agreement was found among platforms characterized by the same field strength. Median CV_{inter}% is reported in Fig. 5 and for both field strength is lower than 2%. For both field strengths all ROIs had an inter-scanner $CV_{\text{inter}} < 5\%$ except for ROI-13 (0.2 7× 10^{-3} mm^2/s) that exceeded this threshold. No significant differences were found for CV_{inter} between 1.5 and 3 T scanners (p = 0.79). In the reduced range CV_{inter} was lower than 2% for all vials except for ROI-8 on 1.5 T scanners ($CV_{\text{inter}} = 3.28\%$). In the isocenter (ROI-1) inter-scanner coefficient of variation was 0.82% for 1.5 T scanners and 1.12% for 3 T scanners.

Overall median CVs_{inter}% along each imaging plane were higher compared to those obtained from averaged measurements.

ADC short-term reproducibility

BA plots for short-term reproducibility are presented in Fig. 6a for 1.5 T, in Fig. 6b for 3 T scanners and summarized

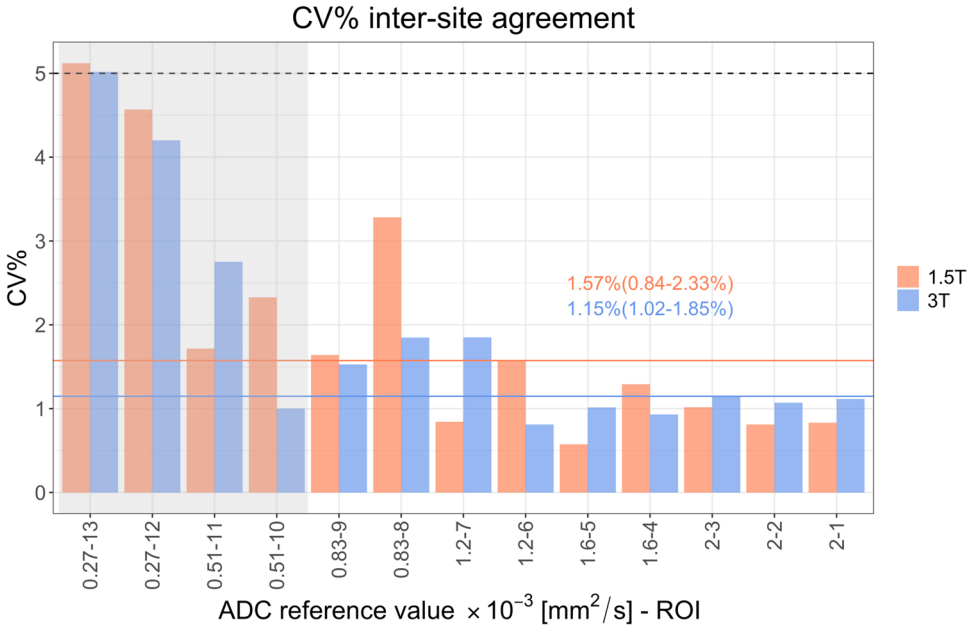


Fig. 5 Inter-scanner coefficient of variation (CV_{inter}%) of average ADC values. Median CV_{inter}% with its lower and upper quartile for each direction and field strength. The area shaded in gray corresponds to the values outside the renal ADC physiologic range. Solid horizontal lines represent CV_{inter}% median values for each field strength (orange for 1.5 T and blue for 3 T) while the dashed horizontal line represents the 5% limit

in Table 4. Pooled results for both field strengths are reported in Fig. 6c. Both single-scanner and pooled BA plots showed a near-zero bias for all sites and field strengths. Limits of agreements were within 10% of bias between the first and second scan for all single scanner BA. Scanners 2, 3, 4, 5, and 6 for 1.5 T and scanners 8, 9, 11 and 12 for 3 T had limits of agreement within 5% of bias. BA outliers were prevalent in the lower ADC range ($< 0.83 \times 10^{-3}$ mm^2/s), no outlier exceeded 10% bias.

Limits of agreement of pooled results were lower than 5% and few outliers were present for average ADC $\leq 0.8\,3\times 10^{-3}$ mm^2/s. All outliers came from measurements performed on vendor A.

BA plots along each imaging plane are reported in Supplementary Fig. S9. Both for 1.5 T and 3 T near-zero bias was verified for all directions and larger limits of agreement were found across axial/sagittal compared to coronal scan plane. Scanner 1, 4 and 10 along axial and sagittal positioning and Scanner 7 only for axial imaging plane had limits of agreement that exceeded 10%.

For average precision error field strength and vendor were tested in a GLMM both through their interaction and as independent predictors (Fig. 4e). The model identified vendor as the only significant independent predictor. Post-hoc tests showed that vendor A had a significantly higher precision error compared to vendor C ($p = 5.08 \times 10^{-3}$).

In the reduced range average precision error had no significant independent predictor thus no significant differences were found between 1.5 and 3 T scanners and among vendors.

Precision error from individual imaging planes was also studied with a GLMM. A significant interaction was found between vendor and field only for coronal acquisitions ($p = 0.01$, Table S5). Along coronal and axial orientation significant differences were found across different vendors (Fig. 4c). No significant differences were found between 1.5 T scanners and 3 T scanners except for vendor A along the coronal plane ($p = 0.03$). In the reduced range no significant differences were found between 1.5 and 3 T scanners and vendors with the only exception of vendor A having a higher precision error than vendor C along axial orientation ($p = 2.30 \times 10^{-3}$) (Fig. S7). Results of precision error comparison across directions for each scanner are reported in Supplementary Fig. S10. Overall, most of the differences were found between coronal direction and axial/sagittal showing that coronal orientation had either a significantly lower precision error or no significant difference from the other two directions. Corresponding ICC results are reported in the Supplementary material (Table S7).

Discussion

In this study we performed a multi-site and multi-vendor assessment of ADC measurements in the three orthogonal imaging planes on 12 scanners evenly distributed by field strength, using the commercially available QIBA diffusion phantom with the vendor-specific standardized NIST protocols and scanners set-ups typical of abdominal studies. In addition, we assessed both intra- and inter-scanner reproducibility of ADC measurements, investigating the effect of several factors such as field strength, vendor, imaging plane orientation, and vials PVP concentration.

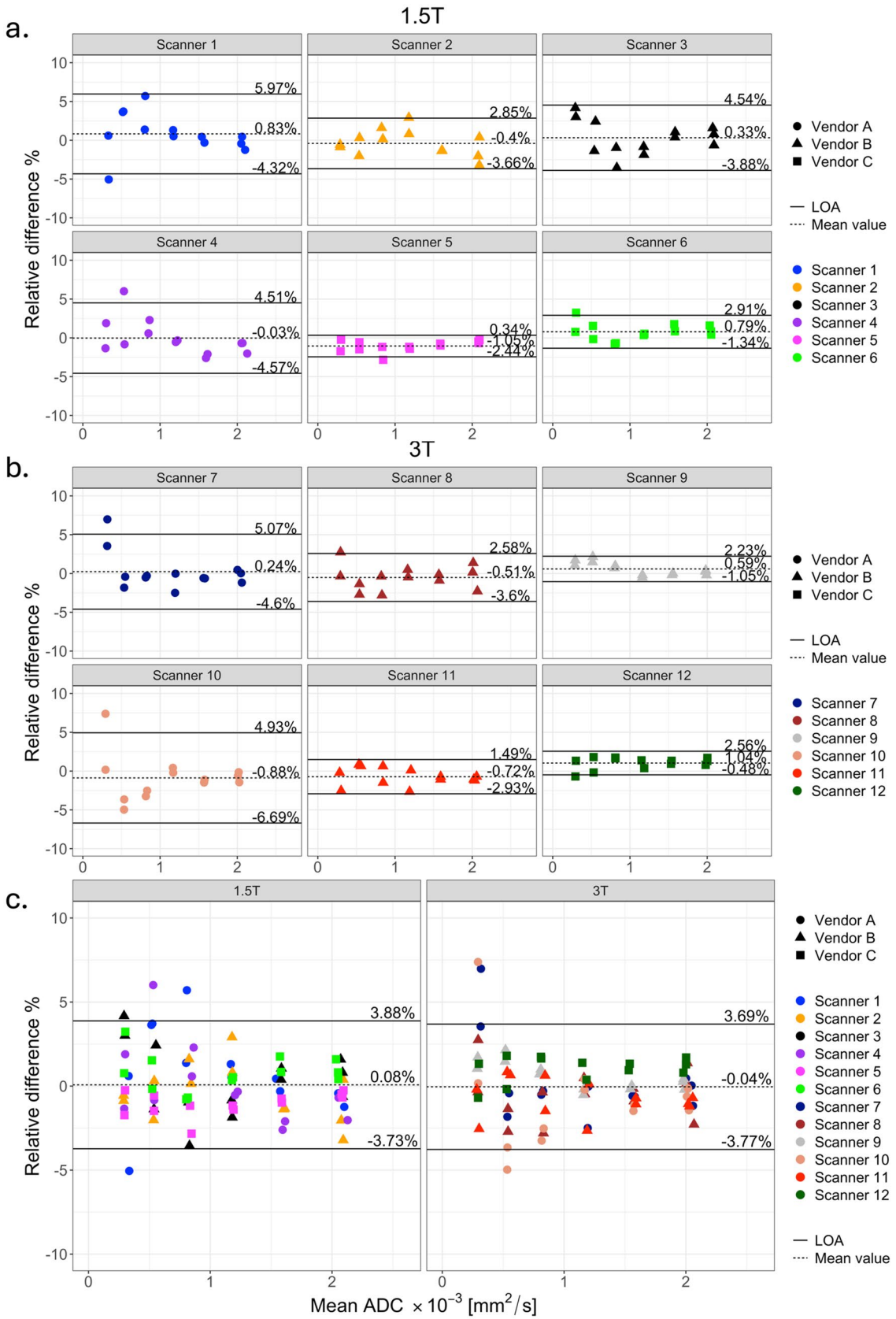
Overall, in each site and across each imaging plane, excellent linear correlation was found between temperature rescaled ADC measurements and NIST reference values, in accordance to previous studies in literature [8, 36]. Accuracy error ranged from 0 to 10% in the renal physiologically relevant range while higher values were found for vials with higher PVP concentrations (and thus lower ADCs). A similar trend was observed for the intra-session coefficient of variation, which exceeded 2% only for 1.5 T outside of the renal physiologic range. As highlighted by previous studies [5, 10, 14, 36, 37] and by QIBA recommendations [6], this may be attributed to several factors such as eddy currents, susceptibility distortions and gradient non-linearities away from the isocenter. The effect of gradient non linearities (GNL) was evident in the spatially dependent bias observed between the pure water vial in the isocenter and inner/outer circle. Results were compatible with GNL bias estimated on water phantoms found in literature. Tao et al. [38] reported a 7% increase of the outer vial compared to reference values, for the same type of phantom in axial positioning. Similarly, Wang et al. [39] found that GNL bias was negligible for a 5-cm offset, reached up to 6% for an offset of 10 cm in the anterior–posterior direction (AP) and for a 15-cm offset it increased to 9.7%, 12.6%, and 9.2% in the right-left (RL), anterior–posterior (AP), and superior-inferior (SI) directions. In the same study, Wang et al. estimated that GNL effect accounted for a significant difference of 5–7% in median value for in vivo renal imaging.

Another factor that may have contributed to accuracy error and reduced repeatability is the insufficient SNR associated with the lower signal from higher PVP concentrations (shorter T2). Indeed, for all scanners, we found evidence of significant correlation between both accuracy error and intra-session variation with SNR. Furthermore, highest accuracy error, highest CV_{intra}% and lowest SNR were found for the same scanner (Scanner 1). Accuracy error and CV_{intra}% in the isocenter for all 3 T scanners, independently on imaging plane position, were within QIBA claims [6] (respectively < 3.6% and < 0.5%) and well supported by other studies [8, 14, 36, 37, 40, 41], despite the use of different phantoms and protocols.

In contrast, only a few of the 1.5 T scanners met accuracy and repeatability QIBA requirements and mainly only along the coronal imaging plane. While some studies [11, 42, 43] also show ADC results not meeting requirements for 1.5 T scanners, either in terms of accuracy or short-term repeatability, others [8, 15, 43] find excellent compliance to QIBA terms. One of the reasons for this difference could be that most phantom studies on 1.5 T were performed at 0 °C thus avoiding potential sources of bias from imperfect temperature estimation and effect of low SNR at high b-values. However, for the purpose of this study and its application to future renal ADC studies, scanner assessment at room temperature was required to cover the renal physiologically relevant range. In addition, most of our 1.5 T scanners did not meet QIBA requirement on SNR > 50. Lastly, as reported by Hoult et al. [44], differences in coil set up compared to the QIBA recommendations (head coil), could have an impact on SNR due to differences in filling factors. Overall, in the isocenter, 94% of ADC values were within the broader range of 5% established by literature [30, 37, 41]. Median accuracy over the whole range was lower than 5% for all scanners and imaging positions while median repeatability was below 1.1%. Our study also showed field strength dependency of accuracy error; 1.5 T scanners compared to 3 T scanners resulted in significantly higher accuracy error for coronal, axial and the overall average. No significant differences were found between different vendors.

Good short-term reproducibility was found for all scanners with limits of agreement within 10%. Precision error in the isocenter for all scanners was compatible with QIBA requirements. We also found that precision error had no significant dependence on field strength ($p = 0.12$) while differences were found among vendors, with vendor A showing the lowest reproducibility. This result may have been caused by using two different body coil set-ups (24 and 48 channels) between the first and second scan session in Scanner 1. Although we acquired an isotropic phantom, significant differences in imaging plane orientation on the same scanner were found for accuracy error, repeatability and precision error. Overall, we found that coronal positioning, which along with axial geometry is the most relevant for renal imaging, was either significantly better or equally good as the other directions. Differences in slice orientation may have been caused by the larger presence of artifacts due to unstable positioning of the phantom when acquired along axial/sagittal plane, geometric distortions in the upper part of the central slice for air bubbles [45] and increased in-plane signal inhomogeneity.

To the best of our knowledge, there are only a few studies [10, 15, 36] that have compared ADC results across imaging plane orientations. However, this is the first study to perform this characterization using an abdominal receive coil. In this study we also found excellent inter-scanner agreement for 1.5 T scanners and 3 T scanners. Results from inter-scanner CV_{inter}% are in line with previous 3 T and 1.5 T multi-center studies: Kooreman [43] observed a median CV across 1.5 T scanners of 2.2%, Fang et al. [14] found a median CV across 3 T scanners < 4%, for PVP < 30% Palacios et al. [12] reported a CV below 3.8% and Wang et al. [40] below 3%. In our study, median CV_{inter} was lower than 2% for both field strengths and for PVP < 30% CV_{inter} was lower than 1.8%. No significant differences were found between 1.5 and 3 T results in terms of inter-scanner agreement.


a.
1.5T
Scanner 1
Scanner 2
Scanner 3
Scanner 4
Scanner 5
Scanner 6
5.97%
0.83%
-4.32%
2.85%
-0.4%
-3.66%
4.54%
0.33%
-3.88%
4.51%
-0.03%
-4.57%
0.34%
-1.05%
-2.44%
2.91%
0.79%
-1.34%
Relative difference %
b.
3T
Scanner 7
Scanner 8
Scanner 9
Scanner 10
Scanner 11
Scanner 12
5.07%
0.24%
-4.6%
2.58%
-0.51%
-3.6%
2.23%
0.59%
-1.05%
4.93%
-0.88%
-6.69%
1.49%
-0.72%
-2.93%
2.56%
1.04%
-0.48%
c.
1.5T
3T
3.88%
0.08%
-3.73%
3.69%
-0.04%
-3.77%
Vendor A
Vendor B
Vendor C
LOA
Mean value
Mean ADC × 10⁻³ [mm²/s]

◂**Fig. 6** Bland–Altman plots of relative difference% between two separate scans average ADC measurements for single scanners at both field strengths (**a**, **b**) and for pooled measurements (**c**). Horizonal dashed line represents the mean relative difference% while horizonal solid lines represent the upper and lower limit of agreements. Values for each line are reported on the right of each panel

To the best of our knowledge, there are no previous multi-site and multi-vendor phantom studies that aim at characterizing ADC measurements in terms of field strength, type of vendor and imaging plane orientation using a surface body coil at room temperature. Our results are strengthened by employing the NIST/QIBA phantom and applying vendor-specific QIBA recommended protocols in each institution.

Limitations

Some limitations need to be acknowledged. Even though we had a moderate sample size of 12 scanners from different vendors, results may have been influenced by the type of scanner and software version. A broader representation of scanners from each vendor category could further strengthen our results. Furthermore, no corrections were applied to account for GNL, geometric distortions and eddy currents, while visible artifacts were simply avoided during ROI placement. For some scanners fewer repetitions of DWI measurements were available, especially along axial and sagittal slice orientation, thus affecting repeatability estimation along these plane orientations. Another limitation of this study was the use of a spherical phantom for abdominal imaging-quality assurance. To better replicate the abdominal geometry, future work should focus on commercially available cylindrical diffusion phantoms. In this study, scanner evaluation was performed using the QIBA-suggested DWI phantom protocol. However, as this may differ from renal imaging protocols used in clinical practice, future work should aim to extend this evaluation to renal clinical protocols, for example using typical b-values of renal studies (b-value < 1000 s/mm^2). In addition, future studies should address common challenges in abdominal imaging, such as: B0 and B1 inhomogeneities due to larger field of view resulting in bias offset, typical b-value ranges to minimize contributions from intravoxel incoherent motion (IVIM) and GNL bias assessment as well as correction at off-center locations using large uniform phantoms. Finally, even though this study was set up to perform scanner assessment for future multi-center kidney clinical studies, the available range of ADC values at room temperature fully covers the physiologically relevant CKD range but only partially the healthy one. A further extension of the range would have required the aid of a heating device to consistently and uniformly increase the temperature of the phantom in all sites, however, aside from technical challenges, comparison with NIST traceable reference values would have been hindered since these are only provided up to 26 °C.

Conclusions

In conclusion, these results from phantom ADC measurements support the feasibility of implementing scanner evaluation using the QIBA standardization process for diffusion measurements in renal studies and contribute to the existing knowledge on standardized multi-site phantom studies at ambient temperature. Despite the use of abdomen coils and room temperature measurements, all quantitative parameters were within literature findings. We found high accuracy, repeatability and precision within the physiologically relevant ADC range for kidneys. Excellent inter-scanner agreement was also found both among 1.5 T and 3 T scanners, supporting possible future renal multi-site studies involving 1.5 T alongside 3 T scanners. Future research studies

Table 4 Bland–Altman plots statistics for each scanner and pooled data, along each imaging plane direction and for average results, reported as mean value of bias% with lower and upper limits of agreement

Scanner	Coronal [%]	Axial [%]	Sagittal [%]	$\overline{ADC}$[%]
1	− 2.25 (–8.35, 3.84)	4.44 (− 11.40, 20.28)	− 0.67 (− 12.89, 11.56)	0.83 (− 4.32, 5.97)
2	1.07 (− 1.37, 3.50)	− 1.74 (− 7.40, 3.90)	− 0.61 (− 6.59, 5.38)	− 0.40 (− 3.66, 2.85)
3	0.98 (− 5.61, 7.56)	− 0.59 (− 3.84, 2.67)	0.53 (− 4.38, 5.44)	0.33 (− 3.88, 4.54)
4	0.69 (− 3.77, 5.14)	− 3.80 (− 20.68, 13.08)	2.37 (− 14.33, 19.07)	− 0.03 (− 4.57, 4.51)
5	− 0.34 (− 2.00, 1.32)	− 0.49 (− 2.61, 1.62)	− 2.40 (− 7.65, 2.86)	− 1.05 (− 2.44, 0.34)
6	0.84 (− 0.77, 2.46)	1.69 (− 0.52, 3.91)	− 0.26 (− 5.34, 4.82)	0.79 (− 1.34, 2.91)
7	− 0.81 (− 3.14, 1.52)	0.87 (− 13.56, 15.29)	0.65 (1.74, 3.03)	0.24 (− 4.60, 5.07)
8	− 0.91 (− 7.52, 5.71)	− 1.09 (− 5.81, 3.63)	0.37 (− 6.04, 6.79)	− 0.51 (− 3.60, 2.58)
9	− 1.15 (− 3.59, 1.29)	3.09 (− 0.44, 6.63)	− 0.29 (− 1.99, 1.40)	0.59 (− 1.05, 2.23)
10	− 0.56 (− 2.65, 1.54)	− 1.16 (− 10.13, 7.81)	− 1.08 (− 10.78, 8.62)	− 0.88 (− 6.69, 4.93)
11	− 1.96 (− 4.88, 0.97)	− 2.19 (− 9.30, 4.91)	1.85 (12.28, − 8.59)	− 0.72 (− 2.93, 1.49)
12	1.51 (− 0.44, 3.48)	− 0.14 (− 4.92, 4.64)	1.72 (− 0.32, 3.76)	1.04 (− 0.48, 2.56)
All 1.5 T	0.16 (− 4.61, 4.94)	− 0.08 (− 10.94, 10.77)	− 0.17 (− 9.79, 9.44)	0.08 (− 3.73, 3.88)
All 3 T	− 0.66 (− 4.61, 3.30)	− 0.11 (− 8.67, 8.44)	0.53 (− 6.12, 7.17)	− 0.04 (− 3.77, 3.69)

focused on clinical MRI protocol validation, first in phantoms and then in healthy volunteers and patients, are needed to enable reliable multi-center MRI clinical studies and ultimately foster the translation of MRI biomarkers to clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10334-025-01256-0.

Author contributions Pasini: Acquisition of data, analysis and interpretation of data, drafting of the manuscript, critical revision. Ringgaard: Study and concept design, critical revision. Vendelboe: Acquisition of data, critical revision. Garcia-Ruiz: Acquisition of data, critical revision. Strittmatter: Acquisition of data, critical revision. Villa: Acquisition of data, critical revision. Raj: Acquisition of data, critical revision. Echeverria-Chasco: Acquisition of data, critical revisionBozzetto Acquisition of data, critical revision. Brambilla: Acquisition of data, critical revision. Aastrup: Acquisition of data, critical revision. Hansen: Acquisition of data, critical revision. Pierotti: Acquisition of data, critical revision. Renzulli: Acquisition of datacritical revision. Francis: Study and concept design, critical revision. Zoellner: Study and concept design, critical revision. Laustsen: Study and concept design, critical revision. Fernandez-Seara: Study and concept design, critical revision. Caroli: Study and concept design, drafting of the manuscript, critical revision.

Funding This study was supported by the Italian Ministry of Health, Gobierno de Navarra, German Federal Ministry of Education and Research (BMBF, grant number 01KU2102), and Innovation Fund Denmark (IFD), under the frame of ERA PerMed (RESPECT project, n. ERAPerMed-2020-326). Dr G. Villa received a scholarship from "Aiuti per la Ricerca sulle Malattie Rare" (A.R.M.R) Foundation.

Data availability Data are available in Zenodo (https://doi.org/10.5281/zenodo.13982980) and will be shared upon reasonable request.

Declarations

Conflict of interest All authors have nothing to declare.

Ethical standards The study does not involve human subjects.

References

1. Le Bihan D, Iima M (2015) Diffusion magnetic resonance imaging: what water tells us about biological tissues. PLoS Biol 13:e1002203
2. Caroli A (2022) Diffusion-weighted magnetic resonance imaging: clinical potential and applications. JCM 11:3339
3. Führes T (2024) Feature-guided deep learning reduces signal loss and increases lesion CNR in diffusion-weighted imaging of the liver. Z Med Phys 34:258–269
4. Caroli A, Schneider M, Friedli I, Ljimani A, De Seigneux S, Boor P, Gullapudi L, Kazmi I, Mendichovszky IA, Notohamiprodjo M, Selby NM, Thoeny HC, Grenier N, Vallée J-P (2018) Diffusion-weighted magnetic resonance imaging to assess diffuse renal pathology: a systematic review and statement paper. Nephrol Dial Transplant 33:ii29–ii40
5. Paquier Z, Chao S-L, Bregni G, Sanchez AV, Guiot T, Dhont J, Gulyban A, Levillain H, Sclafani F, Reynaert N, Bali MA (2022) Pre-trial quality assurance of diffusion-weighted MRI for radiomic analysis and the role of harmonisation. Physica Med 103:138–146
6. Boss MA, Malyarenko D, Partridge S, Obuchowski N, Shukla-Dave A, Winfield JM, Fuller CD, Miller K, Mishra V, Ohliger M, Wilmes LJ, Attariwala R, Andrews T, deSouza NM, Margolis DJ, Chenevert TL (2024) The QIBA profile for diffusion-weighted MRI: apparent diffusion coefficient as a quantitative imaging biomarker. Radiology 313:e233055
7. Boss MA, Chenevert TL, Waterton JC, Morris DM, Ragheb H, Jackson A, deSouza N, Collins DJ, van Beers BE, Garteiser P, Doblas S, Russek SE, Keenan KE, Jackson EF, Zahlmann G (2014) Temperature-controlled Isotropic Diffusion Phantom with Wide Range of Apparent Diffusion Coefficients for Multicenter Assessment of Scanner Repeatability and Reproducibility. Proc 22nd Int Soc Magnet Reson Med. 4505
8. Choi SJ, Kim KW, Ko Y, Cho YC, Jang JS, Ahn H, Kim DW, Kim MY (2024) Whole process of standardization of diffusion-weighted imaging: phantom validation and clinical application according to the QIBA profile. Diagnostics 14:583
9. Shukla-Dave A, Obuchowski NA, Chenevert TL, Jambawalikar S, Schwartz LH, Malyarenko D, Huang W, Noworolski SM, Young RJ, Shiroishi MS, Kim H, Coolens C, Laue H, Chung C, Rosen M, Boss M, Jackson EF (2019) Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. Magn Reson Imaging. https://doi.org/10.1002/jmri.26518
10. Yung JP, Ding Y, Hwang K-P, Cardenas CE, Ai H, Fuller CD, Stafford RJ (2020) Quantitative evaluation of apparent diffusion coefficient in a large multi-unit institution using the QIBA diffusion phantom. https://doi.org/10.1101/2020.09.09.20191403
11. Van Houdt PJ, Kallehauge JF, Tanderup K, Nout R, Zaletelj M, Tadic T, Van Kesteren ZJ, Van Den Berg CAT, Georg D, Côté J-C, Levesque IR, Swamidas J, Malinen E, Telliskivi S, Brynolfsson P, Mahmood F, Van Der Heide UA (2020) Phantom-based quality assurance for multicenter quantitative MRI in locally advanced cervical cancer. Radiother Oncol 153:114–121
12. Palacios EM, Martin AJ, Boss MA, Ezekiel F, Chang YS, Yuh EL, Vassar MJ, Schnyer DM, MacDonald CL, Crawford KL, Irimia A, Toga AW, Mukherjee P (2017) Toward precision and reproducibility of diffusion tensor imaging: a multicenter diffusion phantom and traveling volunteer study. AJNR Am J Neuroradiol 38:537–545
13. (2023) QIBA profile conformance testing DWI MR, Supplement 1, Version 20231102
14. Fang LK, Keenan KE, Carl M, Ojeda-Fournier H, Rodríguez-Soto AE, Rakow-Penner RA (2023) Apparent diffusion coefficient reproducibility across 3 T scanners in a breast diffusion phantom. Magn Reson Imaging 57:812–823
15. Aylward B (2022) Calibration of ADC values for quantitative MRI. Master's thesis, University of Glasgow
16. Zhang JL, Sigmund EE, Chandarana H, Rusinek H, Chen Q, Vivier P-H, Taouli B, Lee VS (2010) Variability of renal apparent diffusion coefficients: limitations of the monoexponential model for diffusion quantification. Radiology 254:783–792

17. Namimoto T, Yamashita Y, Mitsuzaki K, Nakayama Y, Tang Y, Takahashi M (1999) Measurement of the apparent diffusion coefficient in diffuse renal disease by diffusion-weighted echo-planar MR imaging. J Magn Reson Imaging 9:832–837
18. Suo S-T, Cao M-Q, Ding Y-Z, Yao Q-Y, Wu G-Y, Xu J-R (2014) Apparent diffusion coefficient measurements of bilateral kidneys at 3 T MRI: effects of age, gender, and laterality in healthy adults. Clin Radiol 69:e491–e496
19. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M (1988) Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology 168:497–505
20. Xu X, Fang W, Ling H, Chai W, Chen K (2010) Diffusion-weighted MR imaging of kidneys in patients with chronic kidney disease: initial study. Eur Radiol 20:978–983
21. Yalçin-Şafak K, Ayyildiz M, Ünel SY, Umarusman-Tanju N, Akça A, Baysal T (2016) The relationship of ADC values of renal parenchyma with CKD stage and serum creatinine levels. Eur J Radiol Open 3:8–11
22. CaliberMRI Model 128 Diffusion Phantom Spec Sheet.
23. Russek SE, Keenan K, Stupic K, Rentz N, Boss M, Coakley KJ, Koepke A, Stoffer C (2023) Magnetic resonance imaging biomarker calibration service: NMR measurement of isotropic water diffusion coefficient. https://doi.org/10.6028/NIST.SP.250-100
24. Keenan KE, Stupic KF, Russek SE, Mirowski E (2020) MRI-visible liquid crystal thermometer. Magn Reson Med 84:1552–1563
25. Bane O, Hectors SJ, Wagner M, Arlinghaus LL, Aryal MP, Cao Y, Chenevert TL, Fennessy F, Huang W, Hylton NM, Kalpathy-Cramer J, Keenan KE, Malyarenko DI, Mulkern RV, Newitt DC, Russek SE, Stupic KF, Tudorica A, Wilmes LJ, Yankeelov TE, Yen Y, Boss MA, Taouli B (2018) Accuracy, repeatability, and interplatform reproducibility of T_1 quantification methods used for DCE-MRI: results from a multicenter phantom study. Magn Reson Med 79:2564–2575
26. Stupic KF, Ainslie M, Boss MA, Charles C, Dienstfrey AM, Evelhoch JL, Finn P, Gimbutas Z, Gunter JL, Hill DLG, Jack CR, Jackson EF, Karaulanov T, Keenan KE, Liu G, Martin MN, Prasad PV, Rentz NS, Yuan C, Russek SE (2021) A standard system phantom for magnetic resonance imaging. Magn Reson Med 86:1194–1211
27. Chenevert TL, Malyarenko DI, Newitt D, Li X, Jayatilake M, Tudorica A, Fedorov A, Kikinis R, Liu TT, Muzi M, Oborski MJ, Laymon CM, Li X, Thomas Y, Jayashree K-C, Mountz JM, Kinahan PE, Rubin DL, Fennessy F, Huang W, Hylton N, Ross BD (2014) Errors in quantitative image analysis due to platform-dependent image scaling. Transl Oncol 7:65–71
28. Dietrich O, Raya JG, Reeder SB, Reiser MF, Schoenberg SO (2007) Measurement of signal-to-noise ratios in MR images: Influence of multichannel coils, parallel imaging, and reconstruction filters. Magn Reson Imaging 26:375–385
29. Jackson E, Bronskill M, Drost D, Och J, Pooley R, Sobol W, Clarke G (2010) Acceptance testing and quality assurance procedures for magnetic resonance imaging facilities. https://doi.org/10.37206/101
30. Fedeli L, Belli G, Ciccarone A, Coniglio A, Esposito M, Giannelli M, Mazzoni LN, Nocetti L, Sghedoni R, Tarducci R, Altabella L, Belligotti E, Benelli M, Betti M, Caivano R, Carni' M, Chiappiniello A, Cimolai S, Cretti F, Fulcheri C, Gasperi C, Giacometti M, Levrero F, Lizio D, Maieron M, Marzi S, Mascaro L, Mazzocchi S, Meliado' G, Morzenti S, Noferini L, Oberhofer N, Quattrocchi MG, Ricci A, Taddeucci A, Tenori L, Luchinat C, Gobbi G, Gori C, Busoni S (2018) Dependence of apparent diffusion coefficient measurement on diffusion gradient direction and spatial position—a quality assurance intercomparison study of forty-four scanners for quantitative diffusion-weighted imaging. Physica Med 55:135–141
31. Statton BK, Smith J, Finnegan ME, Koerzdoerfer G, Quest RA, Grech-Sollars M (2022) Temperature dependence, accuracy, and repeatability of T_1 and T_2 relaxation times for the ISMRM/NIST system phantom measured using MR fingerprinting. Mag Reson Med 87:1446–1460
32. R Core Team (2022) R: A language and environment for statistical computing
33. Gałecki A, Burzykowski T (2013) Linear mixed-effects models using R: a step-by-step approach. https://doi.org/10.1007/978-1-4614-3900-4
34. Walker L, Curry M, Nayak A, Lange N, Pierpaoli C, the Brain Development Cooperative Group (2013) A framework for the analysis of phantom data in multicenter diffusion tensor imaging studies: framework for Multicenter DTI Analysis. Hum Brain Mapp 34:2439–2454
35. Pang Y, Malyarenko DI, Amouzandeh G, Barberi E, Cole M, Vom Endt A, Peeters J, Tan ET, Chenevert TL (2021) Empirical validation of gradient field models for an accurate ADC measured on clinical 3T MR systems in body oncologic applications. Physica Med 86:113–120
36. Carr ME, Keenan KE, Rai R, Boss MA, Metcalfe P, Walker A, Holloway L (2022) Conformance of a 3T radiotherapy MRI scanner to the QIBA diffusion profile. Med Phys 49:4508–4517
37. Malyarenko D, Galbán CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, Ross BD, Chenevert TL (2013) Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. Magn Reson Imaging 37:1238–1246
38. Tao AT, Shu Y, Tan ET, Trzasko JD, Tao S, Reid RD, Weavers PT, Huston J 3rd, Bernstein MA (2018) Improving apparent diffusion coefficient accuracy on a compact 3T MRI scanner using gradient nonlinearity correction. J Magn Reson Imaging 48:1498–1507
39. Wang J, Ma C, Yang P, Wang Z, Chen Y, Bian Y, Shao C, Lu J (2023) Diffusion-weighted imaging of the abdomen: correction for gradient nonlinearity bias in apparent diffusion coefficient. J Magn Reson Imaging 58:223–231
40. Wang Y, Tadimalla S, Rai R, Goodwin J, Foster S, Liney G, Holloway L, Haworth A (2021) Quantitative MRI: Defining repeatability, reproducibility and accuracy for prostate cancer imaging biomarker development. Magn Reson Imaging 77:169–179
41. Chenevert TL, Galbán CJ, Ivancevic MK, Rohrer SE, Londy FJ, Kwee TC, Meyer CR, Johnson TD, Rehemtulla A, Ross BD (2011) Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. Magn Reson Imaging 34:983–987
42. Moreau B, Iannessi A, Hoog C, Beaumont H (2018) How reliable are ADC measurements? A phantom and clinical study of cervical lymph nodes. Eur Radiol 28:3362–3371
43. Kooreman ES, Van Houdt PJ, Nowee ME, Van Pelt VWJ, Tijssen RHN, Paulson ES, Gurney-Champion OJ, Wang J, Koetsveld F, Van Buuren LD, Ter Beek LC, Van Der Heide UA (2019) Feasibility and accuracy of quantitative imaging on a 1.5 T MR-linear accelerator. Radiother Oncol 133:156–162
44. Hoult DI, Richards RE (2011) The signal-to-noise ratio of the nuclear magnetic resonance experiment. 1976. J Magn Reson 213:329–343
45. Lewis B, Guta A, Mackey S, Gach HM, Mutic S, Green O, Kim T (2021) Evaluation of diffusion-weighted MRI and geometric distortion on a 0.35T MR-LINAC at multiple gantry angles. J Applied Clin Med Phys 22:118–125