Magnetic Resonance Materials in Physics, Biology and Medicine (2025) 38:611–627
https://doi.org/10.1007/s10334-025-01260-4

RESEARCH ARTICLE

Multi-center and multi-vendor evaluation study across 1.5 T and 3 T scanners (part 2): T1 and T2 standardization in the ISMRM/NIST MR 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. Zöllner[4,5] · Christoffer Laustsen[2] · Maria A. Fernandez-Seara[3] · Anna Caroli[1]

Received: 12 January 2025 / Revised: 17 April 2025 / Accepted: 27 April 2025 / Published online: 17 May 2025

Abstract

Objective To assess multi-site and multi-vendor accuracy, and intra- and inter-scanner variability of T1 and T2 measurements using the ISMRM/NIST System MRI phantom at room temperature.

Materials and methods T1 and T2 measurements were acquired using standardized NIST protocols on 13 scanners (1.5 T and 3 T) from 3 vendors at 7 sites and compared with reference values at room temperature. Pearson's correlation (r) and accuracy error were used for comparison with reference values, while inter-scanner agreement was assessed using the coefficient of variation (CV%). Short-term reproducibility was evaluated using Bland–Altman plots and precision error. Generalized linear mixed models and post hoc tests ($\alpha = 0.05$) were adopted to compare accuracy and precision across field strengths, vendors, and scanners. T2 measurements were corrected with StimFit toolbox for stimulated echo compensation.

Results T1 and T2 measurements had excellent correlation with reference values at both field strengths. Stimfit significantly improved T2 accuracy in the renal range for 9 of 13 scanners. Short-term reproducibility (limits of agreement < 10%) and inter-scanner agreement were good (median CV < 7%) for both T1 and T2 values. Inter-scanner CV was < 5% in the renal range for both parameters.

Discussion These findings support the need of scanner evaluation processes to ensure reliable T1–T2 measurements in multi-center MRI studies.

Keywords Phantom · MRI · Relaxation times · 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

T1- and T2-mapping are widely used quantitative MR imaging techniques that can offer valuable insights into the physical properties of tissues [1]. The potential of both longitudinal (T1) and transverse (T2) relaxation times as effective quantitative imaging biomarkers has been extensively demonstrated across numerous clinical applications [2–5], including the renal field, as recently emphasized in a review by Wolf et al. [6].

However, extending these techniques to multi-center studies often poses challenges, as comparisons are often hindered by differences in sequence implementation, hardware, and type of vendor [7]. This critical issue was recently highlighted by Boudreau et al. in the 2020 ISMRM/NIST reproducibility challenge [8] and has since been a focus of various multi-center, multi-vendor studies employing the ISMRM/NIST phantom.

In particular Bane et al. [9] performed accuracy, precision, and inter-platform variability T1 measurements in a multi-center study comprising ten scanners across three major vendors and at both field strengths. Results showed field strength, acquisition protocol, and sample influence on the estimation of accuracy, repeatability, and reproducibility. However, the uneven distribution of scanners between the field strengths highlighted the need for validation with a higher number of 1.5 T scanners. This gap was addressed in a subsequent NIST-led study by Keenan et al. [10] which performed a larger multi-site accuracy study with nine 1.5 T scanners from two vendors and eighteen 3 T scanners from three vendors. In this study, no consistent pattern of discrepancies across vendors was found. Additionally, results from the ISMRM/NIST reproducibility challenge reported by Boudreau et al. revealed that among 27 MR scanners (3 T, across three major vendors), inter-scanner variability exceeded intra-scanner variability. This study also stressed the need for more acquisitions to address non-physiological factors contributing to variation in T1 quantitative measurements. All these multi-site studies focused on T1 measurements and were conducted using only one set of vials ($NiCl_2$) from the ISMRM/NIST phantom. In contrast, a recent study by Keenan et al. investigated T1 and T2 accuracy across multiple 0.55 T MRI systems, utilizing both standard protocols and vendor-neutral sequences on all available arrays ($NiCl_2$ and $MnCl_2$). Both this study and the work by Li et al. [11] emphasize the benefit of correcting T2 maps using stimulated echo compensation techniques. In particular, they demonstrated improved results in terms of inter-scanner variability [7] and accuracy [11]. Li et al. also emphasized the importance of including a broader range of scanners, both from the same vendor and across different vendors, to more effectively assess inter-scanner variations.

The aim of this study, conducted 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), is to build upon previous MRI phantom studies and extend available knowledge on the accuracy, and intra- and inter-scanner variability of T1 and T2 measurements across vendors and sites by (i) including both 1.5 T and 3 T scanners from each site, (ii) utilizing an additional vial array in the phantom, which is known to be more temperature-dependent, and (iii) exploring the feasibility of using the StimFit package for T2 value estimation on 1.5 T scanners.

Materials and methods

Phantom

T1 and T2 data were acquired on the ISMRM/NIST System phantom [12] (Model 130 for all scanners but Scanner 11, equipped with Model 106, CaliberMRI, Boulder, CO, USA) as specified in Table 1. The ISMRM/NIST phantom (Fig. 1a) contains two different plates with 14 element arrays of varying $NiCl_2$ and $MnCl_2$ concentrations. The vials locations in each array are shown in Fig. 1c, d, while their corresponding concentrations can be found in Table S1. Reference values at 3 T were provided by NIST Boulder at temperatures of 16, 18, 20, 22, 24, and 26 °C, while 1.5 T reference values at 20 °C were measured by CaliberMRI using NIST methodology. The phantom also contains 10 liquid crystal (LC) MR-readable thermometer vials [13], as illustrated in Fig. 1b, that enable estimation of the phantom temperature in the range between 15 and 24 °C.

Image acquisition

Acquisitions were performed in seven different sites and 13 different scanners providing data from $n = 6$ 1.5 T scanners and $n = 7$ 3 T scanners (Table 1). Scanners were divided among the three main vendors [$n = 4$ from GE Healthcare (Waukesha, WI), $n = 5$ from Siemens Healthineers AG (Forchheim, Germany), and $n = 4$ from Philips Healthcare (Best, the Netherlands)]. For each scanner, acquisitions were performed with the transmit body coil and the available anterior body and spine receive coils. As suggested by the phantom manufacturer, T1 maps were acquired using a 2D fast spin-echo inversion recovery (IR) sequence [12] both on $NiCl_2$ and $MnCl_2$ vials, while T2 maps were acquired only on $MnCl_2$ vials using a multi-echo spin-echo (MESE) sequence [12]. Only for vendor A, T2 maps were acquired with a protocol that differed from the manufacturer's recommendation, allowing the acquisition of all echoes in a single-echo

Table 1 MRI scanner details

Scanner ID	Site	Field strength (T)	Vendor	Vendor ID	Scanner type	Phantom model	Phantom temperature (°C)					
							T1 ($NiCl_2$)		T1 ($MnCl_2$)		T2 ($MnCl_2$)	
							Test	Retest	Test	Retest	Test	Retest
1	IRFMN	1.5	GE	A	Optima 450w	130	20.1	20.6	22.0	20.1	20.1	20.0
2	UNAV	1.5	Siemens	B	MAGNETOM Aera	130	22.0	20.9	21.7	21.5	21.4	21.5
3	UHEI	1.5	Siemens	B	MAGNETOM Aera	130	22.3	22.2	22.3	22.2	22.2	22.3
4	AUH	1.5	GE	A	Optima 450W	130	20.5	20.1	20.6	20.3	21.2	21.0
5	AUH	1.5	Philips	C	Achieva dStream	130	20.5	19.5	20.7	19.5	20.4	21.0
6	UNIBO	1.5	Philips	C	Ingenia	130	20.8	20.5	20.8	20.5	20.1	20.3
7	IRFMN	3	GE	A	Discovery 750W	130	19.5	20.2	20.0	20.4	20.4	20.6
8	UNAV	3	Siemens	B	MAGNETOM Skyra	130	22.9	20.9	22.9	20.9	22.5	21.0
9	UHEI	3	Siemens	B	MAGNETOM Skyra	130	25.0	24.0	25.0	24.0	25.0	24.0
10	AUH	3	GE	A	Discovery 750	130	20.8	20.0	20.8	20.0	20.9	20.7
11	AUH	3	Siemens	B	MAGNETOM Skyra	130	20.5	18.3	20.5	20.0	21.0	20.7
12	UoN	3	Philips	C	Ingenia	106	20.2	20.8	20.2	20.8	20.2	20.8
13	UNIBO	3	Philips	C	Ingenia	130	19.6	19.0	19.5	20.0	19.3	19.5

More details on temperature estimation and uncertainty are reported in the Supplementary Materials

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

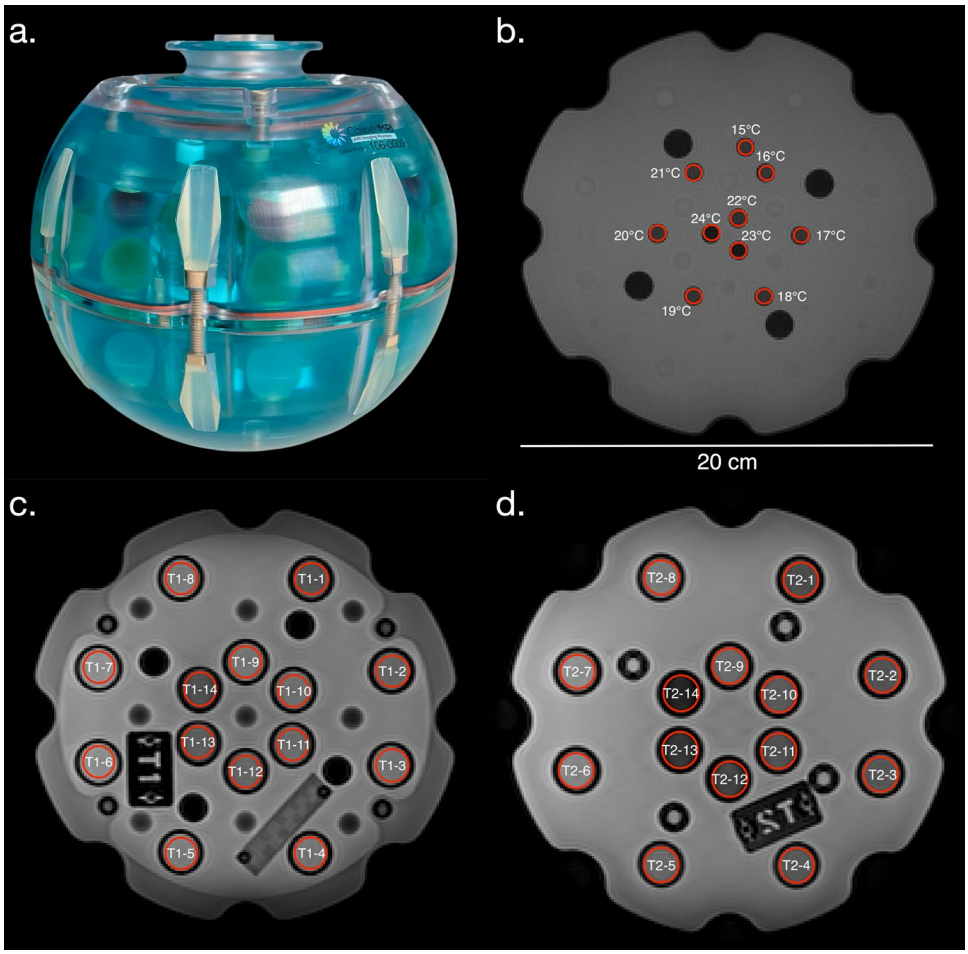


Fig. 1 **(a)** ISMRM/NIST System phantom. Coronal view of the central plate showing regions of interest (ROIs) size and positioning of temperature (**b**), $NiCl_2$ (**c**), and $MnCl_2$ vials (**d**)

train. For Scanner 1, 7, and 10, T2 MESE mapping from UKRIN-MAPS [14] was implemented, while for Scanner 4, we used a Fast Spin Echo (FSE) with 8 echoes available on the scanner. Key parameters for each sequence are reported in Table 2. For each acquisition, a single central slice across the vials was imaged. Following the manufacturer's recommendation, vials T2-1 and T2-5 in Model 130 and T2-2 in Model 106 were excluded from the analysis. The phantom was positioned with its center at the scanner's isocenter, and the plates were aligned with the scanner's principal axes. To minimize movement, due to the phantom spherical shape, cushions were used for stabilization. The exact position of the phantom along the anterior–posterior (A/P) direction depended on the specific setup available for each scanner (see Supplementary Material Fig. S1 for more details on phantom positioning and stabilization).

Each scan was repeated on a different day to measure intra-site reproducibility across test–retest sessions.

As previously described in Part 1 [15], the phantom was placed in the scanner room at least 24 h before each acquisition to allow temperature stabilization. Temperature was measured before and after each session using an isotropic spoiled gradient-echo sequence on the MR-readable thermometer embedded in the phantom [13].

Image processing

Image analysis was centrally performed by a single operator (S.P.), ensuring consistency in the methodology employed. PhantomViewer software [12] was used for image processing. For the central slice, circular regions of interest (ROIs) with an average diameter of 10 mm (~ 82 voxels) were manually placed on each vial while avoiding the edges (Fig. 1b, c). For T2 maps acquired with the UKRIN-MAPS protocol, a larger ROI (13 mm of diameter) was drawn to include at least 35 voxels in each ROI. Mean ROI signal intensity within each ROI was calculated. Adjustments to ROI size and position were made only when visible motion artifacts were present, ensuring accurate signal estimation. For Philips scanners, signal intensity values were appropriately rescaled [16] before calculations. For all scanners, only magnitude data were used in the analyses. Curve fitting was performed in Python using non-linear least-square minimization, with parameters constrained to physically reasonable ranges. For each vial, T1 values were calculated with a 4-parameter model [9, 17] as given by

$$S(TI) = S_0\left|1-(1+\delta)e^{-\frac{TI}{T1}}+\delta e^{-\frac{TR}{T1}}\right|+n,$$

while T2 values were estimated with a 3-parameter mono-exponential model[17]

Table 2 T1 and T2 sequence details

Parameters	T1-IR			T2-SE			
Vendor	A	B	C	A		B	C
Field strength (T)	1.5 T and 3 T	1.5 T and 3 T	1.5 T and 3 T	1.5 T and 3 T		1.5 T and 3 T	1.5 T and 3 T
Acquisition sequence	2D/FSE-IR	2D/TSE-IR	2D/IR-SK	T2 map UKRIN-MAPS	FSE_8echos	2D/se_mc	2D/SE; Fast = None
Orientation	COR	COR	COR	COR		COR	COR
TE (ms)	Min Full	6.9	7	12.9 ms × 10 echoes	10 ms × 8 echoes	10 to 320 ms by 10 ms	11 ms × 16 echoes
TR (ms)	4500	4500	4500	4500	5000	5000	5000
TI (ms)	50, 75, 100, 125, 150, 250, 1000, 1500, 2000, 3000	35, 75, 100, 125, 150, 250, 1000, 1500, 2000, 3000	35, 75, 100, 125, 150, 250, 1000, 1500, 2000, 3000	NA		NA	NA
FOV (mm^2)	250 × 200	250 × 250	250 × 250	250 × 250		250 × 250	250 × 250
# slices	1	1	1	1		1	1
Slice thickness (mm)	6	6	6	4.5	6	6	6
Matrix (FE/PE/SE)	256/192/1	256/192/1	256/252/1	256/192/1	128/128/1	256/192/1	256/192/1
Number of averages	1	1	1	1	1	1	1

FA flip angle, *FOV* field of view, *TR* repetition time, *TE* echo time, *T1-IR* T1 inversion recovery, *T2-SE* T2 spin echo, *FE* frequency encoding, *PE* phase encoding, *SE* slice encoding, *UKRIN-MAPS* UK renal imaging network-MRI acquisition and processing standardization

$$S(TE) = S_0 e^{-\frac{TE}{T2}} + n.$$

In the models, *TI* is the inversion time, *TE* is the echo time, S_0 is the signal amplitude for equilibrium magnetization, δ is the inversion efficiency, *TR* is the fixed repetition time, and *n* is the noise factor. Details on the parameters initial values and constraints that were provided to the fit are reported in Table S2.

T2 values were also estimated using the "StimFit" package provided by UKAT [18] which models the effect of stimulated echoes based on the EPG algorithm [11, 19]. Estimated relative B_1 maps were also derived from StimFit ($0 < B_1 < 1$). Results of StimFit estimation and fit comparisons are reported in Supplementary material (Figs. S2, S3).

Temperature correction

Since both vial arrays, particularly $MnCl_2$ vials, exhibit temperature dependence at both field strengths [12], rescaling to a common reference temperature was fundamental to compare measurements across sessions, scanners, and reference values. Temperature estimation was performed with PhantomViewer software, with ordered circular ROIs (5.5 mm of diameter) placed over the LC MR-readable thermometer vials, as depicted in Fig. 1d. Details on temperature estimation are reported in the Supplementary Material (Fig. S4).

Considering the temperature dependence of the vials [1], temperature adjustments were made using linear rescaling for $MnCl_2$ vials (Figs. S5, S6), as outlined by Statton et al. [10], and quadratic scaling for $NiCl_2$ vials (Fig. S7). In both cases, scaling parameters were derived from the reference values provided in the manual. For 1.5 T scanners, since reference values were available only at 20 °C, temperature adjustment was performed assuming field independent scaling parameters.

Statistical analysis

Statistical analysis was performed using R Studio (version 4.2.1) [20], both in the full range covered by the vials and in the physiologically relevant range for kidneys. $NiCl_2$ vials exhibit T1 values ranging from 22 to 1760 ms at 3 T (20–1723 ms at 1.5 T), while MnCl2 vials show T1 values between 80 and 2480 ms (86–2210 ms at 1.5 T) and T2 values between 5 and 552 ms (7–695 ms at 1.5 T). At the reference temperature of 20 °C both vial arrays cover a broader range than the expected renal physiological one. Based on the values reported by Wolf et al. [6, 21] and on the available reference values from the vials, we selected a physiological range from 690.08 to 1901.28 ms for T1 measurements (685–1741 ms at 1.5 T) and from 44.24 to 267.29 ms for T2 measurements [11] (62–355 ms at 1.5 T). Statistical analysis was conducted using the same approach outlined in Part 1 [15].

For each vial array, agreement between temperature-adjusted measurements and the provided NIST reference at 20 °C was tested with Pearson's correlation coefficient and estimated by computing the relative bias (%). Accuracy error [9] was calculated as follows:

$$AccuracyError\% = 100 \bullet \frac{\left|x_{adjusted} - x_{reference}\right|}{x_{reference}}.$$

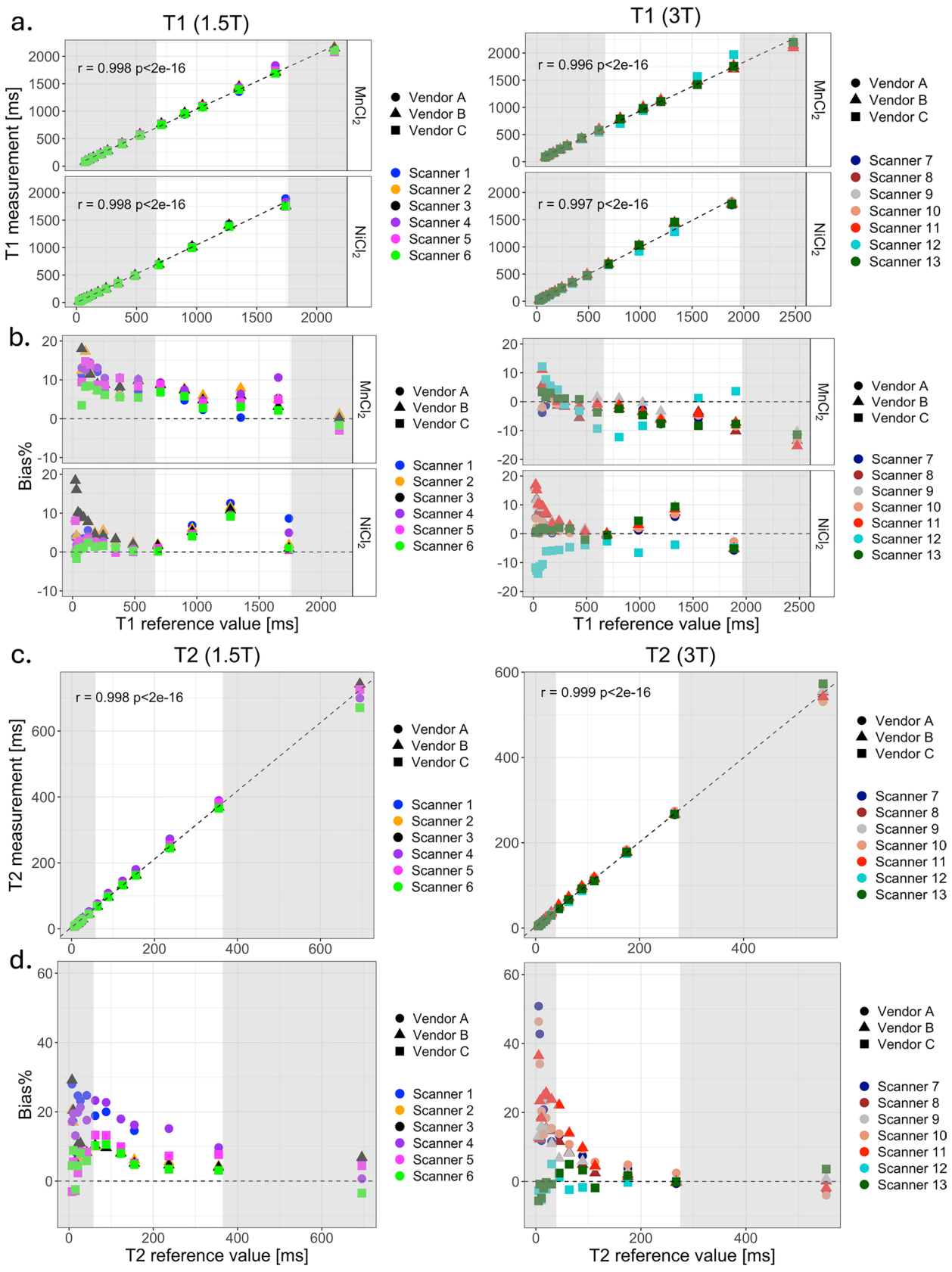
For each scanner and vial array, accuracy error is summarized by its median value along with the lower and upper quartiles. T2 measurements from the first acquisition were compared between the "Stimfit" model and the mono-exponential model, assessing goodness of fit via R^2 and accuracy error for both the full and reduced ranges. Accuracy differences between the two models were analyzed for each scanner using paired sample Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons. The best-fitting model (higher R^2) for each vial was subsequently chosen for the definitive measurements.

Bland–Altman plots were used for visual assessment of short-term reproducibility, while precision error [9] provided a quantitative measure. Precision error was determined as the percentage of the difference between two different scans relative to their average value

$$PrecisionError\% = 100 \bullet \frac{\left|x_{scan1} - x_{scan2}\right|}{Mean\left(x_{scan1}, x_{scan2}\right)}.$$

For each vial array, a generalized linear mixed model (GLMM) was used to compare accuracy error and precision error across field strengths and vendors taking into account repeated measures [9]. For each dataset, field strength and vendor were included as fixed effects of the model both as independent predictors and as two-way interaction. Scanner ID and vial ID were introduced as random effects. The phantom model was introduced both as a fixed effect and as a random effect. In the precision error analysis of T1 measurements, sphere type ($NiCl_2$ or $MnCl_2$) was also modeled as a random effect. Using the "anova" function from the "stats" package in R, model performance metrics were compared to select the best model and the significance of predictors was tested using type 3 *p* values. Bonferroni-corrected post hoc pairwise comparisons were performed with the "emmeans" package. Statistical significance was set at $p < 0.05$.

Inter-scanner agreement was assessed using the inter-scanner coefficient of variation CV_{inter}%, computed as the standard deviation-to-mean ratio for the first session measurements across scanners. Paired sample Wilcoxon signed-rank tests were used for group comparisons. Inter-scanner agreement calculations were also repeated exclusively for


a.
T1 (1.5T)
T1 (3T)
T1 measurement [ms]
r = 0.998 p<2e-16
r = 0.996 p<2e-16
r = 0.998 p<2e-16
r = 0.997 p<2e-16
MnCl2
NiCl2
b.
Bias%
T1 reference value [ms]
c.
T2 (1.5T)
T2 (3T)
T2 measurement [ms]
r = 0.998 p<2e-16
r = 0.999 p<2e-16
d.
Bias%
T2 reference value [ms]
Vendor A
Vendor B
Vendor C
Scanner 1
Scanner 2
Scanner 3
Scanner 4
Scanner 5
Scanner 6
Scanner 7
Scanner 8
Scanner 9
Scanner 10
Scanner 11
Scanner 12
Scanner 13

◂**Fig. 2** Correlation plots of average temperature-rescaled T1 (**a**) and T2 (**c**) values and NIST reference for 1.5 T scanners and 3 T scanners. Correlation lines obtained from a linear interpolation are represented as black dashed lines. Pearson's correlation coefficients with their corresponding p value are displayed in the upper left corner (**a**–**c**) Bias% of temperature corrected T1 (**b**) and T2 (**d**) values plotted with respect to NIST reference for both 1.5 T scanners and 3 T scanners. Horizontal dashed lines represent 0% bias. The area shaded in gray corresponds to the values of T1 and T2 outside of the renal physiological range

scanners utilizing the same phantom model. Intra-class (ICC_{intra}) and inter-class correlation coefficient (ICC_{inter}) were calculated for each array vial and at both field strengths to assess intra-site and inter-site variance contributions [22]. These calculations were performed both on pooled measurements at each field strength and separating across vendors.

Results

T1 results

T1 measurements agreement with NIST reference values

At both field strengths and for both arrays, T1 values estimated with the IR method had an excellent correlation with NIST reference values both at a single-scanner level ($r > 0.990$ for 1.5 T scanners and 3 T scanners) and pooling all scanners together ($r > 0.998$ at 1.5 T and $r > 0.996$ at 3 T). Correlation plots for T1 measurements are reported in Fig. 2a, while bias% is shown in Fig. 2b.

Across all scanners and both vials arrays, bias remained within 20% over the whole T1 range. In the renal physiological range, bias was generally lower 10%, with a few outliers (T1-2 for Scanner 1, 2, and 3, T2-3 for Scanner 4 and T2-7 for Scanner 12). In all cases, bias tended to increase for lower relaxation times.

For the $NiCl_2$ vials, the bias distribution was consistent across field strengths, with values exceeding 10% for T1 < 100 ms and in the reduced range only for vial T1-2 (T1 = 1270 ms at 1.5 T). Among 3 T scanners, only Scanner 12 showed a different bias distribution, characterized by a systematic underestimation across the entire range, with bias values falling below − 10% for T1 < 100 ms.

In contrast, $MnCl_2$ vials showed distinct patterns across field strengths. At 1.5 T, bias was consistently positive across the whole range, except for the highest T1 value (T2-1). At 3 T, most bias results for T1 > 100 ms showed an underestimation of reference values. For the $MnCl_2$ array, bias similarly tended to increase for T1 < 100 ms ranging from 0 and 20% at 1.5 T and from -5 and 15% at 3 T.

T1 accuracy across scanners

Median accuracy error results for both $NiCl_2$ and $MnCl_2$ arrays are reported in Table 3. Values were calculated both across the full and reduced range. In the full range, median accuracy error among all scanners ranged from 1.45 to 10.48% for the $MnCl_2$ array and from 1.05 to 6.53% for the $NiCl_2$ array. While no significant interaction between field strength and vendor was observed in the generalized linear mixed model for the $MnCl_2$ array ($p = 0.27$), the interaction was significant for the $NiCl_2$ array ($p = 0.005$) (Supplementary Table S3). GLMM on $NiCl_2$ spheres showed significant differences across vendors both at 1.5 T and 3 T (Fig. 3a). At both field strengths, accuracy error was significantly higher for vendor B compared to the other vendors. Only at 1.5 T, a significant difference was also observed between vendor A and C, with vendor C having a significantly lower accuracy error. A significantly higher accuracy error was found for 1.5 T scanners compared to 3 T scanners only for vendor A. For the $MnCl_2$ array, accuracy error was significantly higher for 1.5 T compared to 3 T scanners and no significant differences were found across vendors (Fig. 3b).

In the renal physiologically relevant range, median accuracy error ranged from 3.10 to 7.70% for $MnCl_2$ vials and from 2.31 to 7.76% for $NiCl_2$ vials. For both arrays, no differences were found in GLMM results across field strength and vendors when applied on the reduced range (Fig. S8).

T1 inter-scanner agreement

The results of inter-scanner agreement, expressed as CV_{inter}%, are presented in Fig. 4a. Across both 1.5 T and 3 T scanners, excellent agreement was observed, with median CV% values below 4% (1.93% for 1.5 T scanners and 3.48% for 3 T scanners). For T1 values > 100 ms, all vials had a CV% within the 5% threshold. At 1.5 T, only the first two vials (T1 = 22 ms and T1 = 31 ms) exceeded the 5% limit, while across 3 T scanners, all values < 100 ms were above this limit. Paired sample Wilcoxon signed-rank tests comparing CVinter% between 1.5 and 3 T scanners revealed significantly higher CV values for 3 T scanners. However, when the same analysis was restricted to 3 T scanners that performed acquisitions using the same phantom model (removing Scanner 12), median CV for 3 T scanners decreased to 1.69% and no significant differences were observed between 1.5 and 3 T scanners (Fig. S9).

T1 short-term reproducibility

Bland–Altman (BA) plots showing short-term reproducibility results for each scanner and for pooled results are reported in Figs. 5a and 6a for both field strengths. Near-zero bias was observed in both single-scanner and pooled

Table 3 Within scanner accuracy error% expressed in terms of median, first and third quartile

Scanner	Accuracy error %					
	T1		T1 Physiologically relevant range		T2	T2 Physiologically relevant range
	$MnCl_2$	$NiCl_2$	$MnCl_2$	$NiCl_2$		
1	7.32 (4.75–11.32)	3.85 (2.19- 5.37)	4.75 (2.28–5.30)	7.76 (5.55–9.63)	17.92 (8.41–23.19)	16.23 (9.08–18.62)
2	7.84 (6.25–9.67)	5.76 (2.84–8.71)	7.70 (6.03–7.84)	4.02 (1.96 -7.41)	8.21 (6.14–9.81)	7.18 (5.66–9.41)
3	8.72 (5.90–9.83)	6.53 (3.57–10.07)	5.90 (4.84–7.35)	3.43 (1.46–6.65)	8.23 (5.84–10.58)	6.60 (4.76–9.26)
4	10.48 (7.45–12.60)	1.80 (0.98–3.68)	7.45 (6.33–9.32)	4.60 (3.33–6.14)	17.59 (15.15–19.84)	17.09 (15.42–21.52)
5	8.56 (4.92–9.45)	2.14 (0.69–3.23)	4.92 (4.24–5.80)	2.83 (0.89–5.79)	7.27 (4.45–8.51)	8.79 (7.36–12.38)
6	5.57 (3.10–6.83)	1.18 (0.66–1.66)	3.10 (2.75 -5.77)	2.50 (0.74–5.36)	4.78 (3.44–8.00)	6.37 (3.77–9.69)
7	2.65 (1.31–5.94)	1.49 (1.06–2.70)	5.94 (2.75–7.77)	3.63 (1.37–5.82)	11.60 (5.61–20.87)	6.53 (4.27–9.93)
8	3.52 (2.47–5.74)	5.92 (2.85–7.96)	3.52 (3.46–5.75)	3.70 (1.79–5.71)	11.59 (2.49–14.72)	3.70 (2.16–7.52)
9	1.45 (0.92–3.36)	5.98 (2.99–9.39)	3.36 (1.07–4.42)	3.05 (1.86–5.10)	8.11 (4.19–13.77)	5.06 (3.15–6.53)
10	2.43 (0.90–4.78)	1.05 (0.76–2.53)	4.78 (3.65–5.95)	2.31 (1.75–3.68)	13.64 (5.62–18.43)	7.51 (5.05–10.45)
11	3.44 (1.34–6.24)	6.20 (3.19–9.99)	4.17 (3.44–6.24)	4.11 (2.33–5.98)	18.35 (4.52–23.77)	7.14 (2.18–12.95)
12	5.49 (3.40–8.84)	5.98 (4.71–11.16)	6.00 (3.03–9.34)	4.34 (3.54–5.27)	1.69 (0.98–2.52)	1.24 (0.24–1.69)
13	3.34 (1.61–7.55)	1.66 (0.94–2.08)	7.55 (4.65–7.73)	4.82 (3.47–6.21)	1.94 (0.75–3.61)	2.13 (1.73–3.08)

Bland–Altman analyses across all field strengths and sites. Limits of agreement did not exceed 10% for single-scanner plots, while pooled plots showed tighter limits of 5%, with a few outliers not exceeding 10%. Across 1.5 T and 3 T scanners, the largest limits of agreement were found for Scanner 1 and 6. Most outliers were in the low T1 range (T1 < 100 ms).

GLMM results on precision error in the full range showed a significant interaction between vendor and field ($p = 0.0018$, Supplementary Table S3) (Fig. 4c). No significant differences were found across vendors for 1.5 T scanners, while, across 3 T scanners, vendor B showed a significantly lower precision error compared to both vendor A and C. When comparing precision error across vendors, a significantly higher precision error was found for 3 T results on vendor C.

In the renal physiological range, no significant effect was found for both predictors thus showing no differences in precision error across vendors and field strengths (Fig. S10). Median precision error ranged from 0.3 to 2.4% in the full range and from 0.5 to 2.3% in the reduced range.

T1 variance contribution

Results of ICC calculations for each vial are reported in Table S4-S5. For both field strength, overall ICC_{inter} was greater than ICC_{intra}; however, for 3 T measurements, there were some exceptions (T1-1, T1-4, and T2-10).

T2 results

T2 data fitting: mono-exponential vs Stimfit

Stimfit correction was successfully applied to all vials with some exceptions in the range for T2 < 20 ms (vials T2-11, T2-12, T2-13, T2-14 for Scanner 7, vials T2-13, T2-14 for Scanner 1, 4, 10, 11, and 13, and vial T2-14 for Scanner 12). Examples from each scanner of StimFit and mono-exponential model results applied on vial T2-6 are reported in Figs. S2, S3.

Table S6 presents median accuracy errors for all scanners, calculated using both the mono-exponential and StimFit models across both the full and reduced ranges. The resulting p values of paired sample Wilcoxon signed-rank tests comparing the accuracy errors between the two models are also included in the table. In the full range, median accuracy error calculated with the mono-exponential model ranges from 2.16% for Scanner 13 to 24.85% for Scanner 11; for StimFit results, it ranges from 1.39% for Scanner 12 to 22.72% for Scanner 1. Significant differences in accuracy error were found for Scanner 2 and 3. Within the physiologically relevant range for kidneys, Scanner 12 exhibited the lowest accuracy error, with values of 2.53% using the mono-exponential model and 1.24% with the StimFit correction. In the same range, the highest accuracy errors were recorded for Scanner 11 (25.75%) using the mono-exponential model and Scanner 4 (17.09%) with the StimFit correction. Results from the Wilcoxon paired test shows significant

differences in accuracy error were found for Scanner 2, 3, 4, 6, 7, 8, 9, 10, and 11. Both in the full and reduced range when using the StimFit model, median accuracy error was reduced with the exceptions and Scanner 1 and 13 only in the full range. Specifically, a median accuracy reduction of 16.12%, 15.14%, and 18.06% compared to the mono-exponential model was found for Scanner 8, 9, and 11 in the reduced range. Notably, these scanners also had the lowest average estimated B1 field with nominal flip angle of 0.74, 0.75, and 0.64.

Across all scanners, goodness of fit (R^2) was consistently higher for the StimFit model compared to the mono-exponential one in the range for T2 > 20 ms. Among 1.5 T scanners, there were a few exceptions coming from Scanner 1 at T2-10 and Scanner 4 at T2-3. Furthermore, Scanner 5 showed no improvement in the fit when using the StimFit correction across the whole vials range. Final results were measured using the best fit for each scanner.

T2 measurements agreement with NIST reference values

Excellent correlation was found between T2 temperature-adjusted measurements and reference values both at a single-scanner level ($r > 0.993$ for 1.5 T and $r > 0.999$ for 3 T scanners) and pooling all scanners characterized by the same field strength ($r > 0.998$ at 1.5 T and $r > 0.999$ at 3 T). Correlation plots for T2 measurements are reported in Fig. 2c, while bias% is shown in Fig. 2d.

Among 1.5 T scanners (Fig. 2d) bias remained within 30% across the whole range exhibiting a consistent overestimation of reference values. Exceptions are observed at the extremes of the vial range (T2-1 and T2-13) for Scanners 5 and 6 from vendor C. Within the physiologically relevant range, scanners from vendor A exhibited higher bias%, exceeding 15% for T2 values under 150 ms, while vendors B and C had bias < 15% within the same range. For vendor B, bias tended to increase over 10% for low relaxation times T2 < 15 ms.

For 3 T, scanners from vendor A and B displayed similar bias distributions with increasing values, up to 50%, for decreasing relaxation times. In contrast, scanners from vendor C had bias% within 10% over the whole range with negative values mainly in the low relaxation time range for T2 < 20 ms.

T2 accuracy across scanners

Median accuracy error results are reported in Table 3. In the full range, median accuracy error ranged from 4.78 to 17.92% for the 1.5 T scanners and from 1.69 to 18.35% for the 3 T scanners. Median accuracy error was larger than 10% for Scanner 1 and 4 among 1.5 T scanners and Scanner 7, 8, 10, and 11 among 3 T scanners. GLMM analysis on T2 accuracy error showed vendor-field significant interaction effect ($p = 3.2e-06$, Supplementary Table S3) when considering both vial ranges (Fig. 3d). For 1.5 T scanners, accuracy error was significantly higher for vendor A compared to vendors B and C, with no significant differences observed between vendors B and C. Among 3 T scanners, vendor C had a significantly lower accuracy error than vendors A and B (Fig. 3d). GLMM applied across vendors showed significant differences between 1.5 and 3 T scanners only for vendor C with 3 T scanners having a significantly lower accuracy error.

In the renal physiological range, median accuracy error ranged from 6.37 to 17.09% for 1.5 T scanners and from 1.24 to 7.51% for 3 T scanners. Vendor comparisons yielded results consistent with those observed in the extended range. When comparing across field strengths, scanners from both vendor A and vendor C had significantly lower accuracy errors at 3 T, while no differences were found for vendor B.

T2 inter-scanner agreement

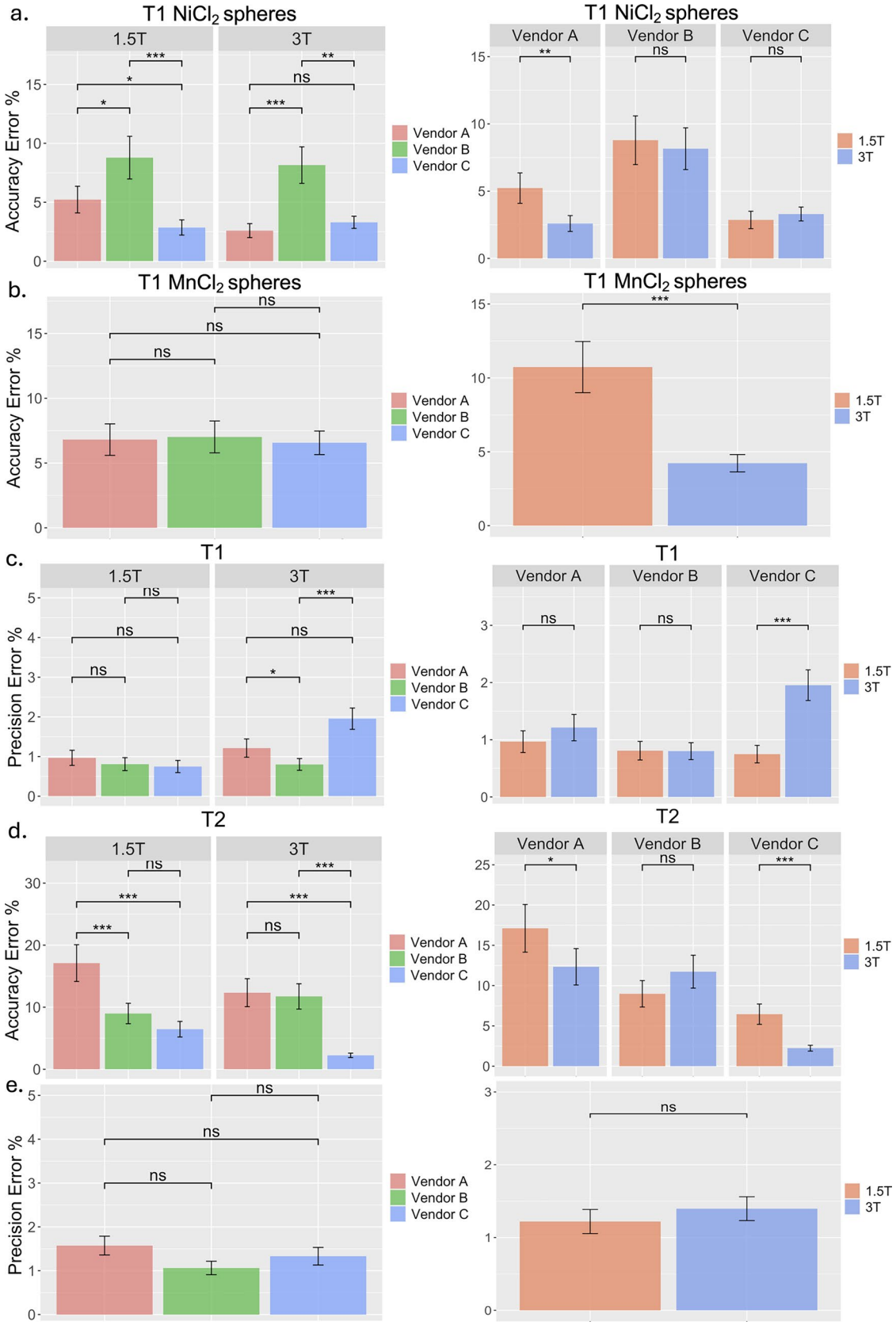
The results of inter-scanner agreement for T2 measurements are reported in Fig. 4b.

For both 1.5 T and 3 T scanners, median CV% values were below 7% (4.95% for 1.5 T scanners and 6.63% for 3 T scanners). For T2 values > 50 ms, thus including the renal range, all vials' CVs% were within 5% showing excellent inter-scanner agreement. For decreasing values of T2, CV% tended to increase, reaching up to 18.89% among 3 T and 14.15% among 1.5 T for the lowest T2 value. Results of paired sample Wilcoxon signed-rank tests comparing CV_{inter}% between 1.5 and 3 T scanners revealed no significant difference between 1.5 and 3 T scanners.

T2 short-term reproducibility

The short-term reproducibility results for each scanner and pooled data are displayed in the Bland–Altman (BA) plots in Figs. 5b and 6b for both field strengths. Both analyses showed near-zero bias across all scanners and field strengths. The limits of agreement for single-scanner data were contained within 5.5%, while pooled data had limits within 5.5%, with only a few outliers above these thresholds. Across 1.5 T and 3 T scanners, the largest limits of agreement were found for Scanner 1, 7, and 9. Most outliers lie in the low T1 range (T2 < 20 ms).

GLMM results on precision error (Fig. 3e) showed that neither vendor nor field strength had a significant effect in both full and reduced range (Supplementary Table S3).


a.
T1 NiCl2 spheres
1.5T
3T
Accuracy Error %
Vendor A
Vendor B
Vendor C
T1 NiCl2 spheres
Vendor A
Vendor B
Vendor C
1.5T
3T
b.
T1 MnCl2 spheres
Accuracy Error %
T1 MnCl2 spheres
c.
T1
Precision Error %
T1
d.
T2
Accuracy Error %
T2
e.
Precision Error %

◂**Fig. 3** Results of GLMM on T1 accuracy error (**a**, **b**), T1 precision error (**c**), T2 accuracy error (**d**), and T2 precision error (**e**). All results are reported as least-square means and standard errors calculated with GLMM. Horizontal lines represent the paired comparisons with their relative significance value (ns = $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

T2 variance contribution

Results of ICC calculations for each vial are reported in Tables S4-S5. For both field strength, overall ICC_{inter} was greater than ICC_{intra}.

Discussion

This study evaluated the accuracy and inter-scanner variation of T1 and T2 measurements across 13 scanners, equally distributed across the three main vendors (GE, Siemens, and Philips) and between field strengths (1.5 T and 3 T). The commercially available ISMRM/NIST System phantom was employed, following vendor-specific standardized NIST protocols and typical abdominal imaging set-ups. Intra-scanner short-term reproducibility was assessed by performing duplicate measurements for each scanner. Additionally, we analyzed the influence of factors, such as field strength, vendor, vial composition, and fit model on the estimation of relaxation times.

The results of this study demonstrated that T1 measurements showed excellent correlation with reference values at both field strengths and for both types of vials. The $NiCl_2$ array showed a common bias trend, with larger values in the range T1 < 100 ms, aligning with findings reported in the previous studies [1, 9, 10, 17]. However, Scanner 12 deviated from this trend, showing an underestimation across the whole range. This was the only scanner that used a different version of the phantom, though it is important to note that the overall behavior, in terms of absolute distance from reference values, was consistent with that of the other scanners. High accuracy error in the low T1 range was also found for the $MnCl_2$ array. Furthermore, in line with the previous studies [9, 17], in this range, we also found a larger variability across scanners, with CV > 5%. These results could be explained by the parameters used in the protocol. Given that the shortest inversion times ranged from 35 to 75 ms (for vendor A 50–75 ms), and T1 values below 100 ms have a null point around $TI_{null} < (100\text{ms} \cdot \ln 2) \approx 70$ ms, in this range, we can expect most of the data along the inversion curve to have reached equilibrium magnetization. As a result, only a limited number of points were available to estimate the curve's longitudinal relaxation time thus contributing to a higher uncertainty in the measurement. Adjusting the selection of inversion times could enhance the accuracy of this estimation. However, since the focus is primarily on the physiologically relevant range, low relaxation times can be disregarded.

Bias% profile was not consistent across spheres and, for the $MnCl_2$ array, also across field strengths. The presence of inconsistent patterns could have been caused by approximate temperature adjustments made at 1.5 T, due to the lack of reference values at different temperatures. This effect is particularly evident for the $MnCl_2$ array, which is more sensitive to temperature changes and field variations compared to the $NiCl_2$ array [10]. Overall, we observed a consistent overestimation of reference values at 1.5 T, while at 3 T measurements overestimated reference values for low T1 values (T1 < 100 ms) but underestimated them in the high T1 range (T1 > 1000 ms). To the best of our knowledge, only one study [7] has performed T1 measurement on the $MnCl_2$ array (T2-11 to T2-3) from the NIST/ISMRM phantom. Similarly to what we found at 1.5 T, at 0.55 T, Keenan et al. [7] found that $MnCl_2$ spheres had a tendency to overestimate reference values with bias up to 15% for long T1 times (> 1200 ms). Results of median accuracy error were in line with the previous findings. Bane et al. [9] conducted a similar study only on $NiCl_2$ spheres reporting median accuracy error ranging from 1.6 to 4.5% across two 1.5 T scanners and from 1.4 to 5.5% across eight 3 T scanners. Expanding on this work, our study included a larger cohort of 1.5 T scanners and extended the analysis to include an additional array of spheres obtaining $NiCl_2$ median accuracy that ranged from 1.18 to 6.53% (5.57–10.48% for $MnCl_2$) across six 1.5 T scanners and from 1.05 to 6.20% (1.45–5.48% for $MnCl_2$) across seven 3 T scanners.

Our study also showed that, while in the full range, some differences were found across vendors both at 1.5 T and 3 T, in the renal physiologically relevant range, field strength and vendor had no significant influence on accuracy error. For $NiCl_2$ accuracy error, significant differences across 3 T vendors were also found by Keenan et al. [7] in their multi-site study, while no differences were found at 1.5 T between vendor A and vendor B. A potential explanation for this discrepancy is the distribution of systems included in the studies (nine 1.5 T scanners with four from vendor A and five from vendor B). In our study, the differences between vendor A and vendor B were close to the threshold of statistical significance ($p = 0.03$) and might become non-significant with a larger sample size of scanners. In contrast, other studies, such as Bane et al. [9], did not identify field strength or vendor as significant predictors; however, their analysis included only two 1.5 T scanners (one from vendor A and one from vendor B).

Excellent inter-scanner agreement (median CV < 4%) and good short-term reproducibility (limits of agreement < 10%) were found across field strengths. These results for $NiCl_2$ sphere type are well supported in the literature. Both Bane

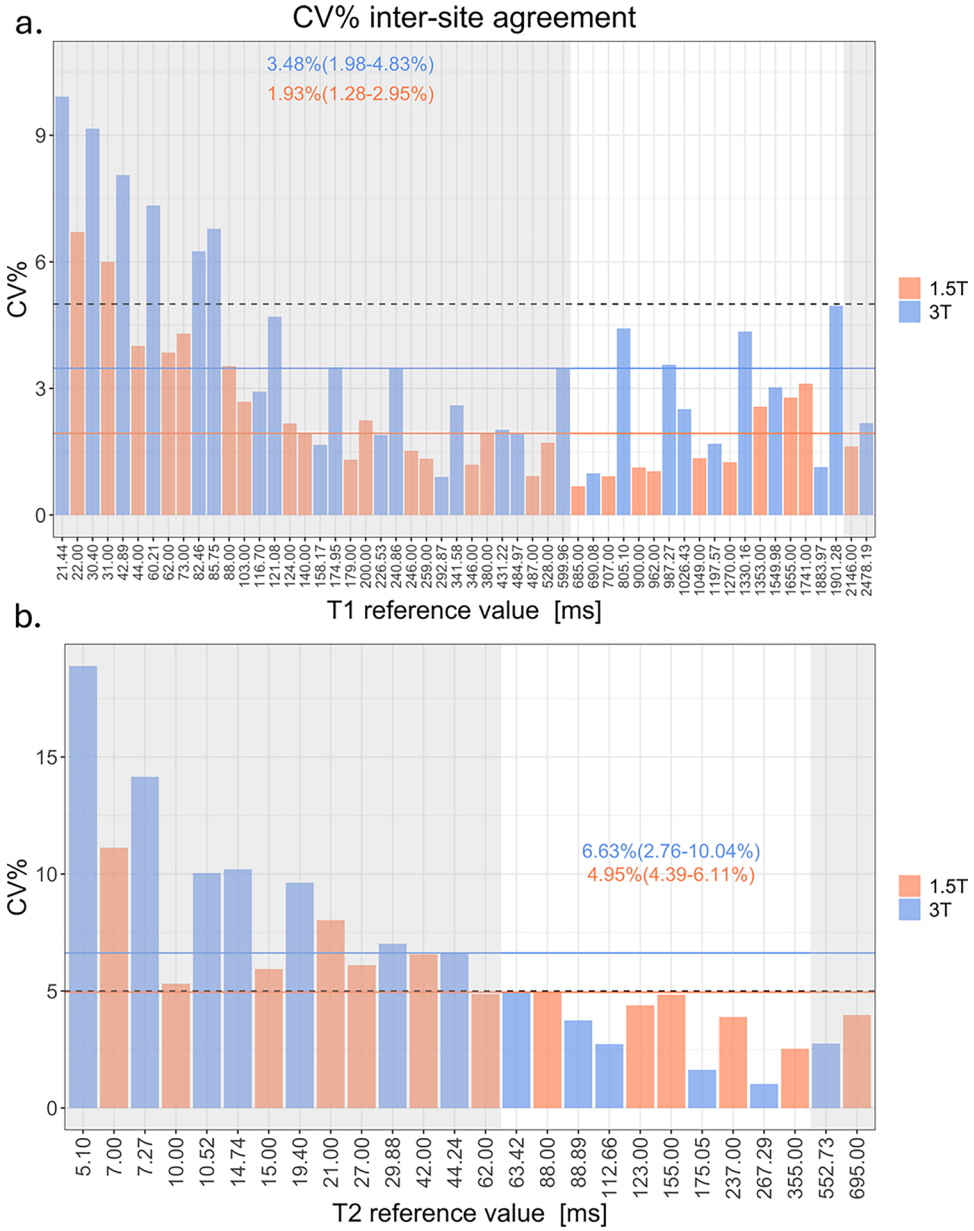


Fig. 4 Inter-scanner coefficient of variation (CV_{inter}%) of T1 values (**a**) and T2 values (**b**). The area shaded in gray corresponds to the values outside of the renal physiological range. Solid horizontal lines represent CV_{inter}% median values for each field strength (orange for 1.5 T and blue for 3 T), while the dashed horizontal line represents the 5% limit. Median CV_{inter}% with its lower and upper quartile for both field strengths are also reported

et al. [9] (RMS CV < 10%) and Van Houdt et al. [23] (median CV across all scanners ≈ 12%) found good inter-platform agreement, with the former also showing no significant differences across field strengths. In our study, no significant differences only emerged when Scanner 12 (phantom Model 130) was included. Median precision error was also in line with values obtained by Bane et al. [9] Precision error in the renal physiological range showed no significant differences across vendors and field strengths. Overall, we found that median inter-scanner variability across vials was larger than intra-scanner variability for both field strengths. Median values for inter-scanner variability decreased if scanners from the same vendors were considered, indicating that most of variability arises from differences between vendors. Only for vendor A, at 3 T, intra-scanner variability was the main source of variance across measurements.

This study also supports the hypothesis that the use of stimulated echo compensation via "StimFit" significantly improves T2 estimation compared to the conventional mono-exponential model by successfully reducing the overestimation expected in multi-echo spin echo due to the presence of stimulated and indirect echoes [24, 25].

This was mainly observed in the physiologically relevant range for kidneys (44.24–267.29 ms at 3 T and 62–355 ms at 1.5 T) at both field strengths. Notably, we found that scanners with the greatest deviations from the ideal flip angle exhibited the largest reduction in overestimation when using Stimfit, highlighting the relation between the severity of

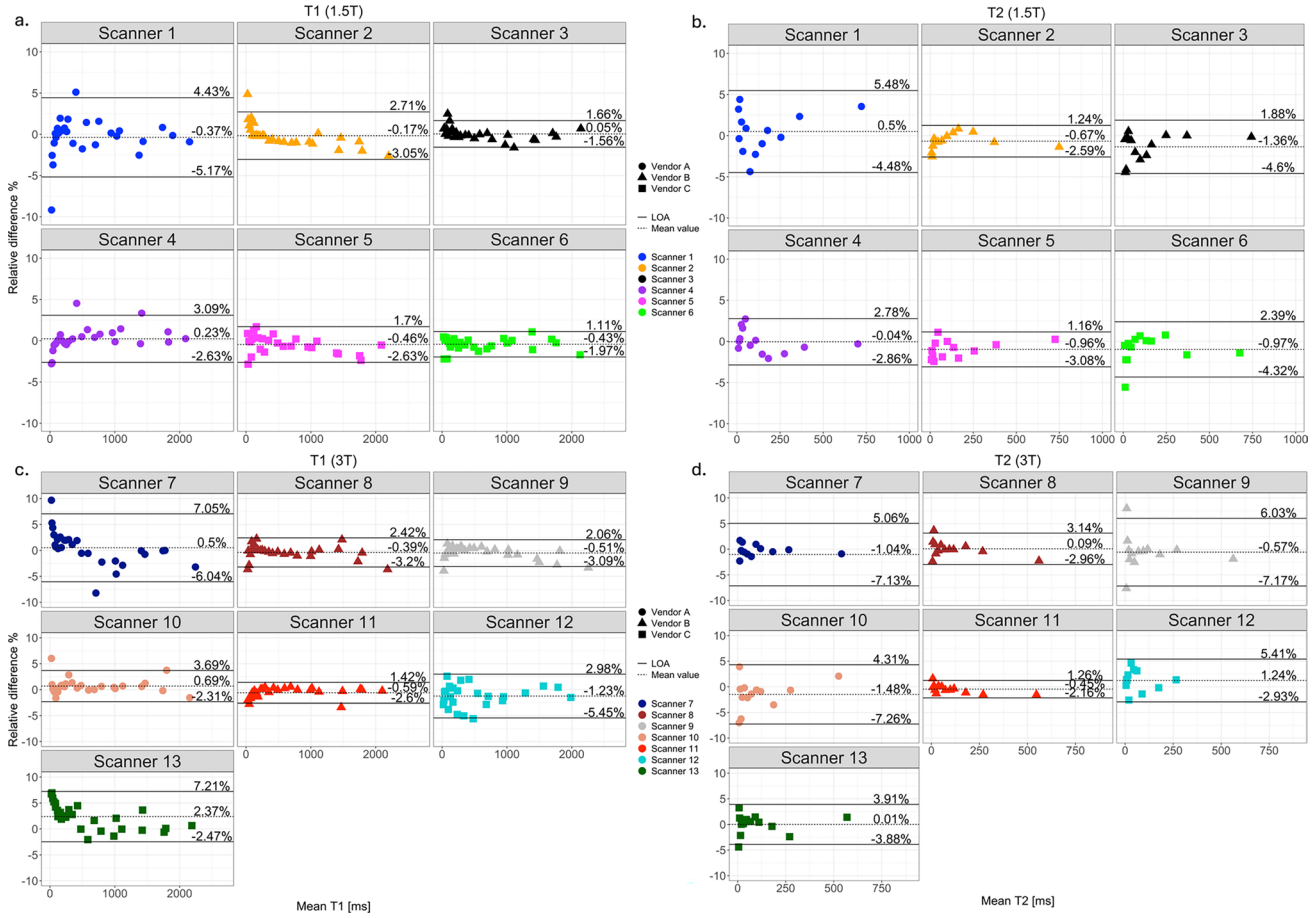


Fig. 5 Bland–Altman plots of relative difference% between T1 (**a**–**c**) and T2 (**b**–**d**) measurements performed in two different sessions for individual scanners at both field strengths. Horizontal dashed line represents the mean relative difference%, while horizontal solid lines represent the upper and lower limit of agreements. Values for each line are reported on the right of each panel

stimulated echoes effect and non-ideal B_1 fields. Across all scanners, no significant improvement was observed only for Scanner 1 and 5 among the 1.5 T group and Scanner 12 and 13 among the 3 T group. Scanner 1 showed a reduction in median accuracy (from 19.81 to 16.45%), with a p value near the significance threshold, suggesting that a larger sample size might yield a significant result. Scanner 5 showed no improvement when using StimFit correction, both in accuracy and goodness of fit, however, this was the only scanner where the first echo was higher than the second echo, suggesting the possibility that the first echo was skipped during acquisition due to specific scanner settings [23]. For Scanner 12 and Scanner 13 (both from vendor C), although the median reduction in accuracy was not statistically significant, the median values remained below the 5% threshold both before and after the application of StimFit correction.

Similar findings were reported by Li et al. [11], who compared the accuracy of mono-exponential fitting and StimFit correction using the NIST protocol on four 3 T scanners: one each from vendors A and C, and two from vendor B. StimFit significantly reduced accuracy errors across all 3 T scanners except for the one from vendor C. This exception was attributed to the scanner's use of composite pulses.

Final T2 measurements showed excellent correlation with reference values at both field strengths. However, distinct bias distributions were observed between the 1.5 and 3 T scanners. All 1.5 T scanners consistently overestimated reference values over the whole range with Scanner 1 and 4 showing the highest median accuracy error, even after StimFit correction in the in vivo range (16.23% and 17.09%). This was also shown in GLMM results, since accuracy error was significantly higher only for vendor A. It should be noticed that for vendor A, the manufacturer-recommended sequences could not be used; instead, measurements were conducted using the T2-mapping sequences available on the scanner. Excluding Scanner 1 and 4, accuracy error for 1.5 T scanners ranged from 6.37 to 8.79% in the physiologically relevant range. Previous studies [23, 26] have reported comparable accuracy errors. One study [26], conducted on four Unity MR-Linac

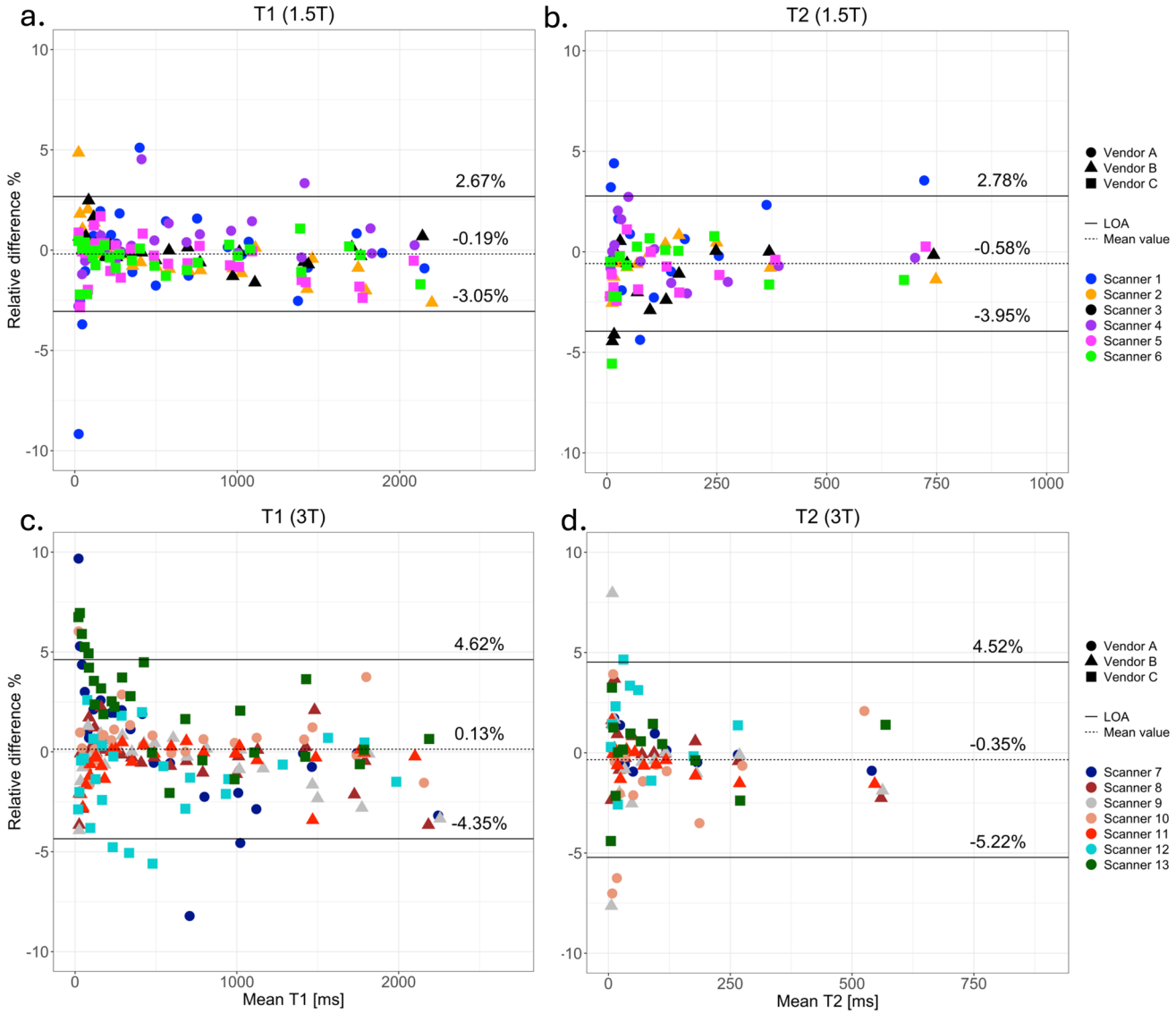


Fig. 6 Bland–Altman plots of relative difference% between T1 (**a**–**c**) and T2 (**b**–**d**) measurements performed in two different sessions pooling measurements at both field strengths. Horizontal dashed line represents the mean relative difference%, while horizontal solid lines represent the upper and lower limit of agreements. Values for each line are reported on the right of each panel

1.5 T systems using the Eurospin phantom, found accuracy errors ranging from 10.4 to 14.1%. Another study [23], which assessed accuracy on 15 scanners (5 at 1.5 T and 10 at 3 T across three vendors), reported accuracy errors for all scanners ranging from 4 to 10%. To the best of our knowledge, this is the first multi-center T2 study at 1.5 T focused on the accuracy and precision of the ISMRM/NIST System Phantom, providing a preliminary estimate of the expected accuracy errors for these measurements. Among 3 T scanners, only those from vendors A and B showed increasing overestimation of reference values for lower T2 measurements. This was also visible in GLMM results, since accuracy error was significantly lower for vendor C. A similar overestimation pattern for low T2 values was observed by Carr et al. [17] only when the first echo was preserved. In contrast, in the same study, they show that removing the first echo resulted in an underestimation of reference values for T2 < 20 ms and a consistent overestimation (median bias of 8.5%) within the in vivo range. In the renal range, all 3 T scanners median accuracy errors were within the 10% range (from 1.24 to 7.51%). These are compatible with the previous results obtained by Li et al. [11] where mean absolute percentage error ranged from 2.9 to 7.3%.

A possible explanation for the differing behaviors between 1.5 and 3 T scanners is the dependence of the Stim-Fit tool on scanner-specific waveforms for excitation and refocusing pulses. Since these waveforms were unavailable

for our scanners, StimFit was implemented using library waveforms provided by the tool. As a result, these corrections should be regarded as approximations.

We found that inter-scanner agreement was excellent in the renal physiological range with all values within 5%. Outside of this range, the CV% increased for lower T2 values especially for T2 < 20 ms, that is also the range where the "Stimfit" correction could not be applied. Overall, higher variability, measurement uncertainty and accuracy errors are expected in this range, as only a limited number of data points are available for fitting the signal decay. Median CV% was compatible with the previous studies [7, 17, 23]. In particular, Carr et al. [17] showed that similar CV% values are obtained even in longitudinal studies on the same scanner.

Reproducibility was good for all scanners with limits of agreements within 10% of relative difference, lowest values of reproducibility were found for low T2 values. Precision error remained below 5% for all vials with median values below 2% across all scanners. A few outliers, reaching up to 10%, were noted for the last two vials thus with the shortest relaxation time (T2-13 and T2-14). No significant differences in precision error were observed across vendors or field strengths, further supporting the high precision of duplicate measurements across all scanners. Overall, we found that the main source of variance across vials came from inter-scanner variability for both field strengths. When restricted to scanners from the same vendor, median inter-scanner variability decreased, indicating that vendor differences were the primary source of variability. For vendor C at both field strengths, intra-scanner variability was larger than inter-scanner variability.

Limitations

While this study provides valuable insights, several limitations should be acknowledged. First, even though the study was well balanced across field strengths ($n = 6$ 1.5 T scanners and $n = 7$ 3 T scanners) and vendors ($n = 4$ for vendor A, $n = 5$ for vendor B and $n = 4$ for vendor C) including more scanner types and software versions for each group could improve the reliability of these findings [27]. Furthermore, while our study focused on reproducibility based on two separate measurements, multiple measurements in time could provide deeper insights. All measurements relied on temperature scaling, which, for 1.5 T scanners, could only be considered approximate, since it was derived from 3 T reference values. Precise temperature scaling for 1.5 T scanners would require NIST-traceable reference values at varying temperatures, similar to those reported by the manufacturer for 3 T scanners. Some limitations were specific to T2 measurements. First, for all scanners from vendor A, the manufacturer-recommended MESE sequence could not be properly implemented, thus requiring the use of alternative sequences available on each scanner. Similar problematics were also found by Li et al. [11]. Additionally, StimFit corrections were successfully applied with some limitations, mainly due to the unavailability of vendor-specific waveforms of excitation and refocusing pulses, especially for 1.5 T scanners, as well as challenges in fitting data within the T2 < 20 ms range. It should be noted that other potential sources of bias in T2 measurements [24], such as diffusion [28, 29], were not studied. Finally, as already outlined in Part 1 of this work, future research should address common challenges in abdominal imaging such as evaluating additional biases due to B0 and B1 field inhomogeneities at larger fields of view, gradient non-linearity, and off-center locations.

Conclusions

To conclude, our study supports the application of the NIST/ISMRM phantom in validating T1 and T2 measurements for renal studies, contributing to the existing knowledge on standardized multi-site phantom studies. We found good reproducibility and excellent inter-scanner agreement for both T1 and T2 measurements at both field strengths especially within the physiological relevant range for kidneys, supporting future multi-site renal studies with both 1.5 T and 3 T scanners. T1 accuracy error was sphere-dependent and, while $NiCl_2$ results were in line with the literature findings, comparisons for $MnCl_2$ were limited by the lack of studies with these vials. Finally, for T2 measurements in the renal range, we found that overestimation was overall reduced with "StimFit" correction even when applied to 1.5 T scanners. These findings are consistent with, and expand upon, a previous study by Li et al. [11] conducted with a smaller number of scanners. To enable reliable multi-center MRI clinical studies and support the translation of MRI biomarkers into clinical practice, future research must focus on validating clinical MRI protocols first in phantoms and then in healthy volunteers and patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10334-025-01260-4.

Acknowledgements 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).

Author contributions Pasini: acquisition of data, analysis and interpretation of data, drafting of the manuscript, and critical revision. Ringgaard: study and concept design, and critical revision. Vendelboe: acquisition of data and critical revision. Garcia-Ruiz: acquisition of data and critical revision. Strittmatter: acquisition of data and critical revision. Villa: acquisition of data and critical revision. Raj: acquisition of data and critical revision. Echeverria-Chasco: acquisition of data

and critical revision. Bozzetto acquisition of data and critical revision. Brambilla: acquisition of data and critical revision. Aastrup: acquisition of data and critical revision. Hansen: acquisition of data and critical revision. Pierotti: acquisition of data and critical revision. Renzulli: acquisition of data and critical revision. Francis: study and concept design, and critical revision. Zoellner: study and concept design, and critical revision. Laustsen: study and concept design, and critical revision. Fernandez-Seara: study and concept design, and critical revision. Caroli: study and concept design, drafting of the manuscript, and 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 No. 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.14568955) 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. Boudreau M, Karakuzu A, Cohen-Adad J, Bozkurt E, Carr M, Castellaro M, Concha L, Doneva M, Dual SA, Ensworth A, Foias A, Fortier V, Gabr RE, Gilbert G, Glide-Hurst CK, Grech-Sollars M, Hu S, Jalnefjord O, Jovicich J, Keskin K, Koken P, Kolokotronis A, Kukran S, Lee NG, Levesque IR, Li B, Ma D, Mädler B, Maforo NG, Near J, Pasaye E, Ramirez-Manzanares A, Statton B, Stehning C, Tambalo S, Tian Y, Wang C, Weiss K, Zakariaei N, Zhang S, Zhao Z, Stikov N, the ISMRM Reproducible Research Study Group and the ISMRM Quantitative MR Study Group (2024) Repeat it without me: crowdsourcing the T_1 mapping common ground via the ISMRM reproducibility challenge. Magn Reson Med 92:1115–1127
2. Taylor AJ, Salerno M, Dharmakumar R, Jerosch-Herold M (2016) T1 mapping: basic techniques and clinical applications. JACC Cardiovasc Imaging 9:67–81
3. Gowland PA, Stevenson VL (2003) T1: the longitudinal relaxation time. Quantitative MRI of the brain. John Wiley & Sons, Ltd, pp 111–141
4. Boulby PA, Rugg-Gunn FJ (2003) T2: The transverse relaxation time. Quantitative MRI of the brain. John Wiley & Sons, Ltd, pp 143–201
5. Press RH, Shu H-KG, Shim H, Mountz JM, Kurland BF, Wahl RL, Jones EF, Hylton NM, Gerstner ER, Nordstrom RJ, Henderson L, Kurdziel KA, Vikram B, Jacobs MA, Holdhoff M, Taylor E, Jaffray DA, Schwartz LH, Mankoff DA, Kinahan PE, Linden HM, Lambin P, Dilling TJ, Rubin DL, Hadjiiski L, Buatti JM (2018) The use of quantitative imaging in radiation oncology: a quantitative imaging network (QIN) perspective. Int J Radiat Oncol Biol Phys 102:1219–1235
6. Wolf M, De Boer A, Sharma K, Boor P, Leiner T, Sunder-Plassmann G, Moser E, Caroli A, Jerome NP (2018) Magnetic resonance imaging T1- and T2-mapping to assess renal structure and function: a systematic review and statement paper. Nephrol Dial Transplant 33:ii41–ii50
7. Keenan KE, Tasdelen B, Javed A, Ramasawmy R, Rizzo R, Marchburn AE, Nayak KS (2025) T1 and T2 measurements across multiple 0.55T MRI systems using open-source vendor-neutral sequences. Magn Reson Med 93:289–300
8. Boudreau M, Karakuzu A, Cohen-Adad J, Bozkurt E, Carr M, Castellaro M, Concha L, Doneva M, Dual S, Ensworth A, Foias A, Fortier V, Gabr R, Gilbert G, Glide-Hurst C, Grech-Sollars M, Hu S, Jalnefjord O, Jovicich J, Stikov N (2023) Results of the ISMRM 2020 joint Reproducible Research & Quantitative MR study groups reproducibility challenge on phantom and human brain T1 mapping. NeuroLibre Reproducible Preprints. https://doi.org/10.55458/neurolibre.00014
9. 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
10. Keenan KE, Gimbutas Z, Dienstfrey A, Stupic KF, Boss MA, Russek SE, Chenevert TL, Prasad PV, Guo J, Reddick WE, Cecil KM, Shukla-Dave A, Aramburu Nunez D, Shridhar Konar A, Liu MZ, Jambawalikar SR, Schwartz LH, Zheng J, Hu P, Jackson EF (2021) Multi-site, multi-platform comparison of MRI T1 measurement using the system phantom. PLoS One 16:e0252966
11. Li H, Daniel AJ, Buchanan CE, Nery F, Morris DM, Li S, Huang Y, Sousa JA, Sourbron S, Mendichovszky IA, Thomas DL, Priest AN, Francis ST (2024) Improvements in between-vendor MRI harmonization of renal T2 mapping using stimulated echo compensation. Magn Reson Imaging. https://doi.org/10.1002/jmri.29282
12. 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
13. Keenan KE, Stupic KF, Russek SE, Mirowski E (2020) MRI-visible liquid crystal thermometer. Magn Reson Med 84:1552–1563
14. Buchanan C, Li H, Morris D, Daniel A, Almeida e Sousa J, Sourbron S, Thomas D, Priest A, Francis S (2022) A Travelling Kidney study using a harmonised multiparametric renal MRI protocol. Conference: The 31st Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM). https://doi.org/10.58530/2022/0482.
15. Pasini S, Ringgaard S, Vendelboe T, Garcia-Ruiz L, Strittmatter A, Villa G, Raj A, Echeverria-Chasco R, Bozzetto M, Brambilla P, Aastrup M, Hansen E, Pierotti L, Renzulli M, Francis S, Zoellner F, Laustsen C, Fernandez-Seara M, Caroli A (2025) 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. Magn Reson Mater Phy. https://doi.org/10.1007/s10334-025-01256-0
16. 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
17. Carr ME, Keenan KE, Rai R, Metcalfe P, Walker A, Holloway L (2021) Determining the longitudinal accuracy and reproducibility of T_1 and T_2 in a 3T MRI scanner. J Appl Clin Med Phys 22:143–150
18. Daniel A, Nery F, Almeida e Sousa J, Buchanan C, Li H, Priest A, Sourbron S, Thomas D, Francis S (2021) UKRIN Kidney Analysis Toolbox (UKAT): A Framework for Harmonized Quantitative Renal MRI Analysis. Conference: The 30th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM)
19. Lebel RM, Wilman AH (2010) Transverse relaxometry with stimulated echo compensation. Magn Reson Med 64:1005–1014
20. R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
21. Wolf M, Kommer S, Fembek S, Dröszler U, Körner T, Berg A, Schmid AI, Moser E, Meyerspeer M (2022) Reproducible phantom for quality assurance in abdominal MRI focussing kidney imaging. Front Phys 10:993241
22. 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
23. 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
24. Fatemi Y, Danyali H, Helfroush MS, Amiri H (2020) Fast T_2 mapping using multi-echo spin-echo MRI: a linear order approach. Magn Reson Med 84:2815–2830
25. Ben-Eliezer N, Sodickson DK, Block KT (2015) Rapid and accurate T_2 mapping from multi–spin-echo data using Bloch-simulation-based reconstruction. Magn Reson Med 73:809–817
26. 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
27. Keenan KE, Gimbutas Z, Dienstfrey A, Stupic KF (2019) Assessing effects of scanner upgrades for clinical studies. J Magn Reson Imaging 50:1948–1954
28. Oakden W, Stanisz GJ (2014) Effects of diffusion on high-resolution quantitative T_2 MRI: effects of diffusion on high-resolution quantitative T_2 MRI. NMR Biomed 27:672–680
29. Carr HY, Purcell EM (1954) Effects of diffusion on free precession in nuclear magnetic resonance experiments. Phys Rev 94:630–638