Magnetic Resonance Imaging 127 (2026) 110578

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

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

Repeatability and reproducibility of maximum diameter measurements of prostate lesions on MRI with repositioning and variation of imaging sequences: A test-retest study

Kevin Sun Zhang [a], Philip Alexander Glemser [a], Christian Jan Oliver Neelsen [a], Markus Wennmann [a], Lukas Thomas Rotkopf [a], Nils Netzer [a,b], Clara Meinzer [a,c], Thomas Hielscher [d], Vivienn Weru [d], Magdalena Görtz [e,f], Albrecht Stenzinger [g], Markus Hohenfellner [e], Heinz-Peter Schlemmer [a,h,j], David Bonekamp [a,h,i,j,*]

[a] Division of Radiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[b] Department of Radiation Oncology, University Hospital Heidelberg, Heidelberg, Germany
[c] Department of Diagnostic and Interventional Radiology with Nuclear Medicine, Heidelberg Thoracic Clinic, University of Heidelberg, Heidelberg, Germany
[d] Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[e] Department of Urology, University Hospital Heidelberg, Heidelberg, Germany
[f] Junior clinical cooperation unit 'Multiparametric Methods for Early Detection of Prostate Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany
[g] Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany
[h] National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany
[i] Heidelberg University Medical School, Heidelberg, Germany
[j] German Cancer Consortium (DKTK), Germany

ARTICLE INFO

Keywords:
Prostate
Magnetic resonance imaging
Reproducibility of results
Observer variation
Dimensional measurement accuracy

ABSTRACT

Objectives: To assess variability of maximum diameter measurements of prostate lesions in MRI assessing patient repositioning, rater and sequence effects.
Methods: Forty-two patients were included retrospectively, who received a clinical bi−/multiparametric prostate MRI examination and agreed to have the T2-weighted (T2WI) and diffusion weighted-imaging (DWI) sequences scanned twice. Maximum diameter measurements of prostate lesions mentioned in the clinical radiologist reports were performed by four readers in multiple reading sessions for determination of inter-sequence (between two DWI sequences), inter-scan (between clinical and additional scan), intra-rater and inter-rater variability. The primary calculated metrics were the repeatability and reproducibility coefficient (RC/RDC), including pooled RC/RDC.
Results: Variability measured by RCs/RDCs was lowest for measurements obtained within the same reading session, with inter-scan RCs up to 5.6 mm/6.5 mm for T2WI/DWI, pooled RCs of 4.8 mm/5.8 mm, respectively, and inter-sequence RDCs of 5.4 mm–5.9 mm, pooled RDC 5.8 mm. Measurements performed in separate reading sessions demonstrated significantly higher variability for both settings in the majority of cases (RCs: up to 10.9 mm/11.7 mm/10.2 mm for T2WI/DWI/inter-sequence, $p \leq 0.002$), pooled RCs/RDCs 9.2 mm–9.9 mm. Measurements necessarily generated in different reading sessions, i.e., intra-rater or inter-rater, demonstrated high variability (RCs/RDCs up to 11.4 mm/11.5 mm for T2WI/DWI).
Conclusions: Prostate lesion measurements demonstrate considerable variability. When measured in one reading session by one rater, lesion diameter differences below the pooled RCs of 4.8 mm, 95 %-CI [3.9, 5.6] for T2WI and 5.8 mm, 95 %-CI [4.7, 7.1] for DWI should not necessarily assumed to be true biological change, as these

Abbreviations: B&A plot, Bland-Altman plot; CI, confidence intervals; COV, coefficient of variation; DWI, diffusion-weighted imaging; LMM, linear mixed model; LOAs, limits of agreement; MLD, maximum lesion diameter; mpMRI, multiparametric magnetic resonance imaging; PCa, prostate cancer; PI-RADS, Prostate Imaging Reporting and Data System; PRECISE, Prostate Cancer Radiological Estimation of Change in Sequential Evaluation; PZ, peripheral zone; R1 / 2 / 3 / 4, reader 1 / 2 / 3 / 4; RC, repeatability coefficient; RDC, reproducibility coefficient; rsEPI, multi-shot echo planar imaging sequence, DWI sequence; SSEPI, single-shot echo planar imaging sequence, DWI sequence; T2WI, T2-weighted imaging; TZ, transition zone.

* Corresponding author at: Department of Radiology (E010), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
E-mail address: d.bonekamp@dkfz-heidelberg.de (D. Bonekamp).

https://doi.org/10.1016/j.mri.2025.110578
Received 27 October 2025; Received in revised form 25 November 2025; Accepted 26 November 2025
Available online 27 November 2025

differences may result from measurement- or repositioning-based variability alone. Caution needs to be taken assessing size changes.

1. Introduction

With increasing use of multiparametric magnetic resonance imaging (mpMRI) of the prostate to detect and stage prostate cancer (PCa) [1–3], there is increasing need to accurately follow mpMRI-detected lesions in patients that do not elect definite treatment. PCa can be sorted into different risk categories, with high-risk tumors demanding immediate treatment, while patients with low-risk and selected intermediate-risk PCs may avoid treatment and elect 'active surveillance' programs, in which their symptoms, laboratory parameters or imaging findings are regularly re-assessed according to American and European guidelines [4,5].

In the Prostate Imaging Reporting and Data System (PI-RADS), mpMRI lesion size measurement is used to differ between PI-RADS category 4 and 5 [3]. While PI-RADS is directed primarily at initial detection and staging of PCa, the Prostate Cancer Radiological Estimation of Change in Sequential Evaluation (PRECISE) scoring system [6,7] provides structured assessment of lesion change over time and is currently undergoing validation [8–11]. Within PRECISE, one of the key metrics for disease regression, stability or progression is lesion size measurement [6]. With its update in March 2024, the PRECISE expert committee proposes a threshold of >50 % volume increase for significant size change [7]. However, the committee did not reach a consensus on a standard approach for lesion measurement, states that single-axis measurement remains important in clinical practice and urges on further research regarding optimal measurement method [7]. PI-RADS currently recommends measurement of peripheral zone (PZ) lesions on the ADC map, and of transition zone (TZ) lesions on T2-weighted imaging (T2WI), while stating that optimal pulse sequence and imaging plane for measurement require further investigation [3]. The PRECISE system prefers measurement on T2-weighted sequences with other sequences aiding in the identification of lesions [7]. In general, there is scarcity of data regarding the variability due to measurement- and positioning-based effects, even for the maximum lesion diameter (MLD) measurement of prostate lesions.

Utilizing a retrospective cohort of patients, who underwent repeat prostate exams on the same day after repositioning, the aim of this study was to determine the minimal threshold for MLD measurements above which a deviation should not be attributed to measurement imprecision alone. This will be achieved by utilizing statistical methods of agreement to quantify variability due to re-positioning, sequence and human rater effects, i.e., intra- and inter-rater effects. Secondary aims are to compare measurement-based variability depending on the prostate zone and the PI-RADS classification of the lesions.

2. Patients and methods

2.1. Study sample

The institutional review board (S-110/2023) approved this retrospective study and waived informed consent. The study was performed in accordance with the Declaration of Helsinki in its current version. Men previously recruited for a repositioning MRI study (04/2017–10/2021) [12,13] were selected for potential inclusion: For the repositioning study, men referred for clinical routine prostate mpMRI were asked for consent to repeat certain MRI sequences (see below), if scanner capacity allowed, e.g. empty time slot until next patient appointment. Inclusion criteria for this retrospective analysis were: (a) successful repeated acquisition of at least one sequence after repositioning in the mentioned study [12] and (b) the mention of at least one focal prostate lesion in the clinical radiologist report. Exclusion criteria were (a) previous focal therapy for prostate cancer, (b) severe imaging artefacts (c) atypical histology or disease (i.e., prostate tuberculosis and leiomyosarcoma).

Test-Retest image acquisition: In brief, patients received an extra MR exam (additional exam), which contained T2-weighted (T2WI) and diffusion-weighted imaging (DWI) sequences identical to the ones of the clinical main exam (clinical exam). The additional exam could be performed before or after the clinical one. Repositioning was performed in-between, i.e. the patient left the scanner, waited for a few minutes and re-entered.

2.2. Imaging and MRI protocol

All patients were imaged on a 3T MRI (Siemens Magnetom Prisma, Siemens Healthineers) with the standard 18-channel body and integrated spine phased-array receiver coils. Lesion assessment was performed using three different sequences: (a) axial T2-weighted turbo spin echo sequence (TR 3710–9370 ms, TE 96–145 ms, slice thickness (ST) 3 mm, in-plane resolution 0.3–0.5 mm); (b) axial singe-shot echo planar imaging (ssEPI, Siemens Healthineers) sequence (b-values: 0,50, 500, 1000, 1500; TR 3300–5700 ms, TE 48–71 ms, ST 3 mm and in-plane resolution 2 mm) and (c) an axial readout-segmented multi-shot echo planar imaging sequence (rsEPI; RESOLVE, Siemens Healthineers); b-values: (0), 50, 1000; TR 4070–6260 ms, TE 45–54 ms, ST 3 mm and in-plane resolution 2 mm). Calculation of the corresponding ADC maps was performed using the scanner's vendor software.

2.3. Image assessment

Two radiology residents with 2 years (R1, KSZ), 1.5 years (R2, CJON), and two board certified radiologists with 10 years (R3, PG) and > 15 years (R4, DP) of experience in prostate mpMRI at the time of the study performed research assessments. Lesion detection was not part of this study, instead R1 created side-by-side hanging protocols with the clinical and additional exams and marked lesions with arrows, which were mentioned in the clinical radiologist reports, using sequence / slice number and pictograms as reference. Readers were instructed to measure each lesion on a single slice in the clinical and additional exam independently on both T2WI images and ADC maps. Scrollable image stacks of the whole prostate were available and the slice to be measured on could be chosen deliberately by the readers to simulate clinical measurements. R1-R3 performed two temporally separate reading sessions while R4 performed only one session, which allowed inter-rater analysis (R3 vs R4). Washout between sessions was 4 weeks minimum. A clinical PACS viewing system - Centricity PACS Radiology RA1000 (GE Healthcare) was used for assessments. Supplementary Fig. S1 gives an example of how measurements were performed.

2.4. Statistical analysis

Comparisons of MLD measurements were made for variability due to (i) inter-scan, (ii) inter-rater, (iii) intra-rater and (iv) inter-sequence (DWI) effects. Mean Bias, Limits of Agreement (LoAs), 95 % confidence intervals (CIs) for LoAs, the corresponding Bland-Altman (B&A) plots, repeatability coefficient (RC) for intra-rater and inter-scan effects, reproducibility coefficient (RDC) for inter-rater and inter-sequence effects, the pooled variants of RCs/RDCs for R1-R3, and coefficient of variation (CoV) were calculated using linear mixed models (LMMs) to account for multiple lesions per patient according to Parker et al. [14], and Zou et al. for the CIs of the LoAs [15]. 95 % confidence intervals for RC/RDC were estimated using bootstrap. Relative RC/RDC was

calculated by division of the coefficient by the respective mean. For CoV, RC and RDC lower values denote better repeatability and reproducibility. The RC and RDC represent the value below which, the absolute difference of two paired measurements would fall within a probability of 95 % [16]. The LMMs were also used to test if the bias was significantly different from zero. Comparison of variances of paired measurements was done according to Bradley and Blackwood [17]. Comparison of variances from different populations, e.g. for lesions of different prostate zones or for pooled data from different PI-RADS categories, were performed according to Levene's test considering the clustered aspect of the data [18]. Holm's method was used to adjust for multiple comparisons [19]. Statistical analysis was performed with R (version 4.0.3; R Foundation for Statistical Computing).

3. Results

3.1. Study sample

Forty-two patients were included for this retrospective analysis. A simplified inclusion diagram and detailed patient information and demographics are given in Fig. 1 and Table 1, respectively. In seven patients, the repeat sequences were performed after contrast agent administration. As diffusion-weighted and T2w images are not substantially affected by the presence of contrast material, especially after the between-exam wait period, exams were pooled for the analysis. The radiologist reports of the 42 patients described 89 lesions, four of which were excluded in consensus due to limited measurability, i.e., v-shaped form, or deemed too subtle and subjective, which resulted in a final lesion count of 85. Initially only the ssEPI sequence was included in the protocol; later the second DWI sequence (rsEPI) was added successively in the clinical and additional scan protocol. Therefore, fewer patients possessed rsEPI scans. The respective, available lesion numbers used for each comparison are given in the respective tables.

3.2. Inter-scan variability: repeatability associated with patient repositioning

When comparing inter-scan MLD measurement performed within two temporally separated reading sessions, absolute RCs ranged between 6.9 mm–7.6 mm/8.0 mm–8.8 mm/10.5 mm–11.7 mm for R1/R2/R3, respectively, corresponding to relative RCs of 54–61 %/70–82 %/71–84 %. Pooled RCs for R1-R3 were 9.9 mm, 95 %-CI [8.1, 11.6], 9.6 mm, 95 %-CI [8.1, 11.3] and 9.7 mm, 95 %-CI [7.4, 11.1] for T2WI/ssEPI/rsEPI, respectively.

However, when using inter-scan measurements generated in the same reading session, variability was significantly smaller with absolute RCs between 4.9 mm–6.5 mm/3.5 mm–6.0 mm/5.3–5.6 mm for R1 ($p < 0.001$ for T2WI-derived measurements) / R2 ($p \leq 0.002$ for all) / R3 ($p < 0.001$ for all), respectively. The respective pooled RCs were 4.8 mm, 95 %-CI [4.0, 5.5] / 5.7 mm, 95 %-CI [4.7, 6.6] / 5.8 mm, 95 %-CI [4.7, 7.1] for T2WI/ssEPI/rsEPI, respectively.

For further details, see Table 2 and Fig. 2. The corresponding B&A plots for R1-R3, shown in Fig. 3 and Supplements S2–3, demonstrate no funnel shape indicating validity of the absolute values for the bias and LoAs.

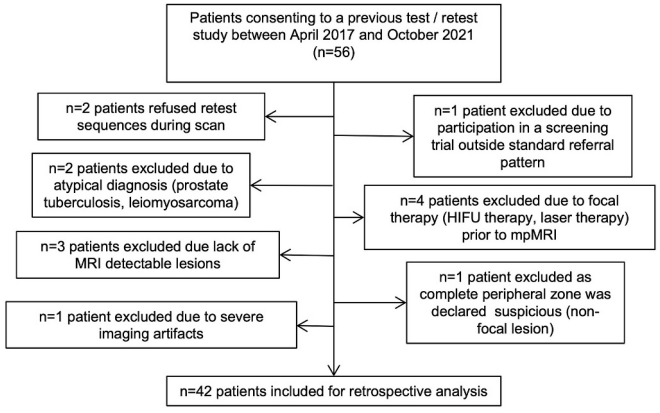

Fig. 1. Inclusion and exclusion diagram.
Abbreviations: mpMRI = multiparametric MRI; HIFU = high-intensity focused ultrasound.

Table 1
Demographic and clinical characteristics of 42 included men.

Cohort	$n = 42$
Age (years)	
Median (IQR)	67 (61–69)
Per-patient maximum ISUP Grade Group (GG) (n (%))	
No Histology Available	7 (17 %)
no PC	13 (31 %)
GG 1	7 (17 %)
GG 2	11(26 %)
GG 3	3 (7 %)
GG 4	0 (0 %)
GG 5	1 (2 %)
PSA (ng/ml) Median (IQR)	7.2 (5.1. – 11.4)
PSA Density Median (IQR)	0.11 (0.08–0.22)
Patient PI-RADS score:	
PI-RADS 1	0 (0 %)
PI-RADS 2	2 (5 %)
PI-RADS 3	16 (38 %)
PI-RADS 4	17 (40 %)
PI-RADS 5	7 (17 %)
Biopsy distribution per patient (n (%))	
Biopsy-naïve	26 (62 %)
Previously biopsied	9 (21 %)
Active surveillance	7 (17 %)
Number of MRI detected lesions per patient	
1	13
2	19
3	7
4	2
5	1
Number of MRI detected lesions	85 (100 %)
PZ	54 (64 %)
TZ with AFS	31 (36 %)
PI-RADS Distribution for the lesions	85 (100 %)
PI-RADS 1	0 (0 %)
PI-RADS 2	5 (6 %)
PI-RADS 3	34 (40 %)
PI-RADS 4	38 (45 %)
PI-RADS 5	8 (9 %)
MRI lesions with sPC (ISUP Grade Group≥2)	17 (100 %)
PZ	14 (82 %)
TZ/AFS	3 (18 %)

Abbreviations: AFS = anterior fibromuscular stroma, IQR = Interquartile Range, ISUP = International Society of Urological Pathology, PSA = prostate specific antigen, MRI = magnetic resonance imaging, PI-RADS = Prostate Imaging Reporting and Data System, sPC = Significant Prostate Cancer, PZ = Peripheral Zone, TZ = Transition Zone.

3.3. Intra-Rater variability: repeatability of one rater's measurements from two reading sessions

Comparing MLD measured on the clinical exams in the two separate reading sessions by each rater, absolute RCs for the different sequences ranged between 7.2 and 9.1 mm/7.2–9.0 mm/10.3–11.5 mm for R1/R2/R3, respectively. The T2WI sequence showed higher variability with an absolute RC of ~9-10 mm for R1-R3. See 'intra-rater' subsection of Fig. 4 and Table 3. Pooled RCs were: 9.5 mm, 95 %-CI [7.9, 11.0] / 9.2 mm, 95 %-CI [7.6, 11.1] / 8.9 mm, 95 %-CI [7.0, 10.6] for T2WI/ssEPI/rsEPI, respectively.

Table 2
Systematic Bias, Coefficient of Variation and Repeatability Coefficient of Maximum Diameter Measurements of prostate lesions in the Inter-Scan setting.

	Rater	Sequence (Number of Lesion)	Bias (LoAs) [mm]	p adj. For significant bias	Mean [mm]	Relative Bias to Mean [%]	CoV [%]	RC (95 % CI) [mm]	Relative RC to Mean [%]
Inter-Scan (separate reading sessions)	R1	T2 (82)	−0.3 (−6.6, 5.9)	1.0	12.5	−2.6	22.2	7.6 (6.0, 9.3)	61
		ssEPI (85)	−0.1 (−5.0, 4.8)	1.0	12.4	−0.7	20.3	6.9 (5.7, 7.9)	56
		rsEPI (52)	0 (−5.4, 5.4)	1.0	13.3	0.1	19.6	7.2 (5.2, 8.8)	54
	R2	T2 (82)	−1.5 (−8.4, 5.4)	0.17	10.7	−13.7	29.8	8.8 (6.9, 10.5)	82
		ssEPI (85)	−1.0 (−7.0, 5.0)	0.42	10.6	−9.6	28.5	8.3 (6.6, 9.6)	78
		rsEPI (52)	−1.0 (−7.3, 5.3)	1.0	11.4	−9.0	25.7	8.0 (5.9, 10.0)	70
	R3	T2 (82)	0.1 (−7.6, 7.8)	1.0	13.6	1.0	28.9	10.9 (8.6, 12.5)	80
		ssEPI (85)	<0.1 (−8.3, 8.3)	1.0	13.9	<0.1	30.6	11.7 (9.9, 13.6)	84
		rsEPI (52)	−0.5 (−8.1, 7.1)	1.0	14.9	−3.3	25.8	10.5 (7.7, 12.7)	71
Inter-Scan (same reading session)	R1	T2 (82)	0 (−3.5, 3.4)	1.0	12.4	0	14	4.9 (3.9, 5.6)	40
		ssEPI (85)	−0.1 (−4.5, 4.3)	1.0	12.4	0	16	5.5 (4.4, 6.5)	44
		rsEPI (52)	−0.8 (−5.4, 3.8)	0.089	13.6	−0.1	17	6.5 (4.7, 7.6)	47
	R2	T2 (82)	0.2 (−2.4, 2.9)	1.0	11.6	2.1	11	3.5 (2.8, 4.2)	31
		ssEPI (85)	0.3 (−4.1, 4.6)	1.0	11.3	2.3	19	6.0 (4.8, 6.9)	53
		rsEPI (52)	0.5 (−3.3, 4.3)	1.0	12.2	4.3	16	5.3 (4.0, 6.4)	44
	R3	T2 (82)	−0.4 (−4.4, 3.6)	1.0	13.3	−3.2	15.3	5.6 (4.4, 6.4)	42
		ssEPI (85)	−0.6 (−5.2, 3.9)	1.0	13.5	−4.6	15.1	5.6 (4.5, 6.7)	42
		rsEPI (52)	−0.6 (−4.5, 3.2)	1.0	14.8	−4.1	13.0	5.3 (3.9, 6.4)	36

Linear mixed models were used to test for significance of bias. * p-values <0.05 were considered significant with Holm's method used to correct for multiple comparisons. Note: Lower CoV, RC/RDC and relative RC/RDC denote better repeatability/reproducibility.
Abbreviations: CI = confidence interval, CoV = coefficient of variation, LoAs = limits of agreement, rsEPI = readout-segmented multi-shot echo planar imaging, ssEPI = single-shot echo planar imaging, R1/2 /3 = Reader 1/2/3, RC = repeatability coefficient.

3.4. Inter-rater variability: reproducibility of measurements between different raters

Inter-rater analysis among either the residents (R1–2) or the board-certified radiologists (R3–4) revealed no significant bias, with the mean bias <1 mm for all sequences ($p \geq 0.42$). Absolute RDCs ranged between 7.3 mm–8.0 mm and 11.0 mm–11.7 mm for the residents and board-certified radiologists, respectively, which corresponds to relative RDCs of 61–67 % and 84–88 %. See 'inter-rater' subsection of Fig. 4 and Table 3.

3.5. Combined inter-rater and inter-scan variability

For assessment of combined inter-rater and inter-scan variability, measurements from a rater on the clinical scan were compared with the measurements from a different rater on the additional scan (R1 vs. R2, and R4 vs. R3). With this approach, combined effects of measurement-based variability of biologically identical lesions are assessed. Absolute RDCs were 10.2 mm/7.8 mm/9.6 mm for T2/ssEPI/rsEPI, respectively, for R1 vs. R2, and 12.1 mm/11.7 mm/11.8 mm, respectively, for R3 vs. R4. See corresponding subsection of Fig. 4 and Table 4.

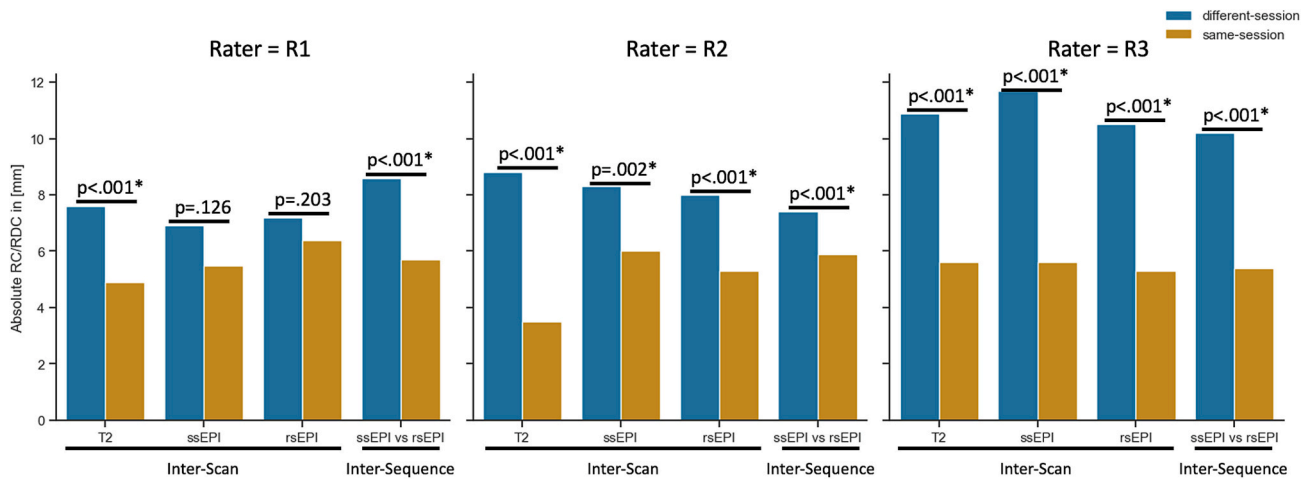

Fig. 2. Absolute repeatability / reproducibility coefficients (RCs/RDCs) of maximum diameter measurements of prostate lesions for the inter-scan repeatability and inter-sequence reproducibility scenarios with comparison of measurements generated in the same and in different reading sessions.
Variance comparison between groups of paired measurements was performed according to Bradley & Blackwood. * *p*-values <0.05 were considered significant with Holm's method used to correct for multiple comparisons.
Abbreviations: rsEPI = diffusion-weighted readout-segmented multi-shot echo planar imaging, ssEPI = diffusion-weighted single-shot echo planar imaging, RC = repeatability coefficient, RDC = reproducibility coefficient.

3.6. Inter-sequence variability: comparison of two different DWI sequences: ssEPI vs. rsEPI

Assessment of inter-sequence variability with measurements performed on ssEPI and rsEPI DWI revealed no significant bias. Absolute RDCs were 7.4 mm/8.6 mm/10.2 mm for R1/R2/R3, respectively, corresponding to relative RDCs of 69–70 % and a pooled RDC of 9.2 mm, 95 %-CI [7.1, 11.3]. Variability was significantly reduced for all raters in same-session measurements, with RDCs between 5.4 mm–5.9 mm ($p <$ 0.001 for all), corresponding to relative RDCs of 39–51 % and a pooled RDC of 5.8 mm, 95 %-CI [4.7, 6.7]. See subsections of Fig. 2 and Table 4. Corresponding B&A plots demonstrate no funnel-shape; see Fig. 3 and Supplements S2–3.

3.7. Variability of sequence-dependent lesion measurements stratified by prostate zone

To investigate repeatability of each sequence depending on the prostate zone, we re-analyzed variability of same-session inter-scan measurements, as this setting demonstrated the smallest RCs so far, however, stratified by prostate zone for this analysis. No significant differences were detected between the zones for all sequences. For graphical illustration, *p*-values according to Levene's test and detailed numbers, see Fig. 5 and Supplements S4.

3.8. Variability of lesion measurements stratified by PI-RADS category

To investigate whether the lesions' PI-RADS category influenced measurement variability, a subanalysis was performed with lesions grouped into PI-RADS≤3 and PI-RADS≥4. Apart from few exceptions, RCs/RDCs were higher for the lower category group in all comparisons indicating higher measurement-based variability. However, differences were not statistically significant. For detailed numbers and p-values according to Levene's test, see Table 5.

3.9. Variability of lesion measurements based on index lesions

As the decision for a change in management is often based on the index lesion, a separate analysis was performed utilizing only index lesions. For this analysis, we used the same-session inter-scan data, which provided the lowest expected variability. If a patient had multiple lesions, the index lesion was defined as the lesion with the highest PI-RADS category or, in case of equal categories, the largest lesion. The tendency towards lower variability with higher PI-RADS category, as reported above, was further supported by this analysis. With analysis of index lesions only, the pooled inter-scan RCs were 3.7 mm, 95 %-CI [3.2, 4.1] / 4.7 mm, 95 %-CI [4.1, 5.2] / 4.8 mm, 95 %-CI [4.1, 5.5] for T2WI/ssEPI/rsEPI, respectively. The pooled inter-sequence RDC (ssEPI/rsEPI) was 4.2 mm, 95 %-CI [3.6, 4.8]. For detailed numbers and individual RCs/RDCs for R1-R3, see Supplements S5.

4. Discussion

Follow-up size measurement is crucial in oncologic imaging for evaluation of disease stability, progression and treatment management. For solid tumors of the body, brain tumors and lymphomas, different response criteria have been developed [20–22]. These classification systems cannot be utilized for evaluation of small intra-prostatic lesions as some are tissue specific [21,22] or designed for larger lesions [20]. The prostate-specific PI-RADS classification does not take into account the temporal change of prostate lesions while the PRECISE expert committee did not reach consensus on optimal measurement method and states that further research is necessary on measurement technique and cut-offs for significant change [3,7,23]. This study was designed to assess variability based on the metrics of bias and repeatability / reproducibility coefficient (RC/RDC), which give the probable maximum difference between two measurements [16].

Maximum lesion diameter (MLD) measurements on prostate MRI demonstrate considerable but similar RCs/RDCs for positioning, rater and inter-sequences effects. Lowest variability was noted for differences after repositioning if measurements were taken in the same reading session with pooled RCs of 4.8 mm and 5.8 mm for T2WI- and diffusion-weighted sequences, respectively. The different imaging sequences did not demonstrate significantly different variability for the two prostate zones. Stratification of lesions according to PI-RADS category (PI-RADS≤3 vs. PI-RADS≥4) demonstrated larger measurement variability for the lower category group, however, differences were not statistically significant. Analysis of the index lesions only, provided lower RCs of 3.7 mm and 4.8 mm for T2WI- and diffusion-weighted sequences, respectively.

Generally, oncologic test/retest studies are scarce, which might be due to high cost and limited scanner time of MRI and radiation exposure

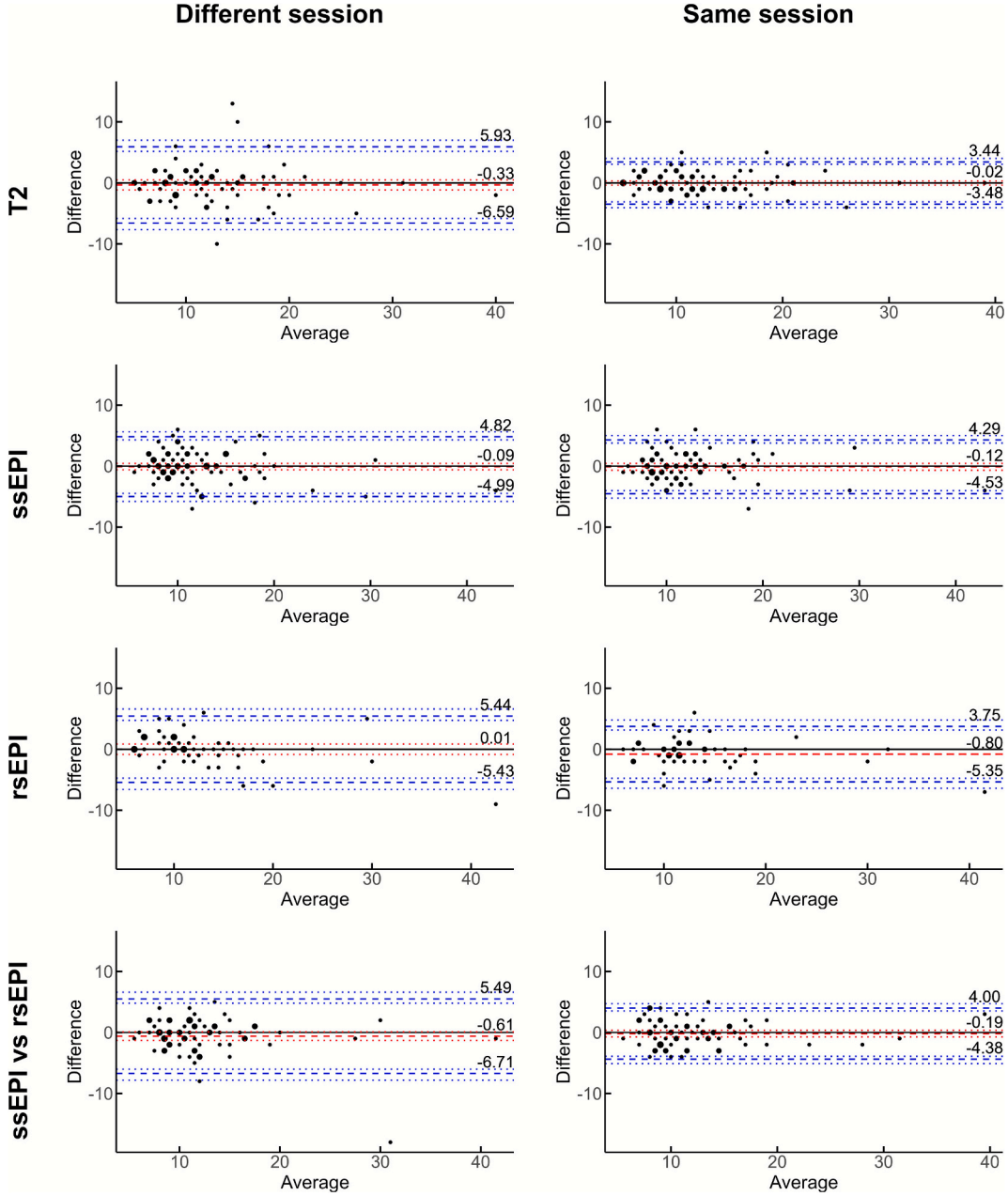

Fig. 3. Bland-Altman plots for Inter-Scan Repeatability and Inter-Sequence Reproducibility of maximum diameter measurements of prostate lesions for T2-weighted and diffusion-weighted (ssEPI and rsEPI) sequences of Reader 1.

The x-axis represents the average of the paired measurements while the y-axis the difference between the measurements. Red dashed lines denote the bias and the red dotted lines the corresponding 95 % CI. The blue dashed lines represent the upper and lower LoAs with the blue dotted lines the corresponding 95 % CI. Values are given in [mm].

Abbreviations: LoAs = limits of agreement, rsEPI = diffusion-weighted readout-segmented multi-shot echo planar imaging, ssEPI = diffusion-weighted single-shot echo planar imaging. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of computed tomography (CT) [24]. Utilizing CT, Zhao et al. and Oxnard et al. reported lower inter-scan variability of lung cancer with LoAs at about ±20 %/±13 % [25,26]. However, comparison to these studies is difficult as tissue-air interfaces in the lung and the higher spatial resolution of CT are expected to improve measurement accuracy compared to prostate MRI, where lesion boundaries can be less distinct. In addition, lesions in these studies are approximately three-times as large as the prostate lesions of our cohort, which may explain increased relative variability in this study, as smaller lesions possess larger relative variability [26,27]. Wennmann et al. investigated inter-scan variability of bone marrow lesions in multiple myeloma patients with mean lesion size of 12.9 mm and LoAs of ±3 mm corresponding to ±23 %, which are similar to the LoAs obtained in this study for T2WI sequences (measurements of same reading session) [24]. In a test/retest study of prostate MRI, Fedorov et al. demonstrated relative RCs of 71 %–112 % for T2w and DWI sequences for lesion volumes; no maximum diameter measurements were given [28]. Further test/retest studies of prostate MRI have been conducted, however, primarily dealing with repeatability and reproducibility of ADC [12,29–32], T1/T2 relaxation times [33] and deep learning performance [34].

Inter- and intra-rater comparison of MLD without test/retest analysis has been performed before [35–37]. However, none of these studies demonstrated absolute agreement metrics like the RC or RDC as shown in this study. In addition, comparison to these studies is limited due to

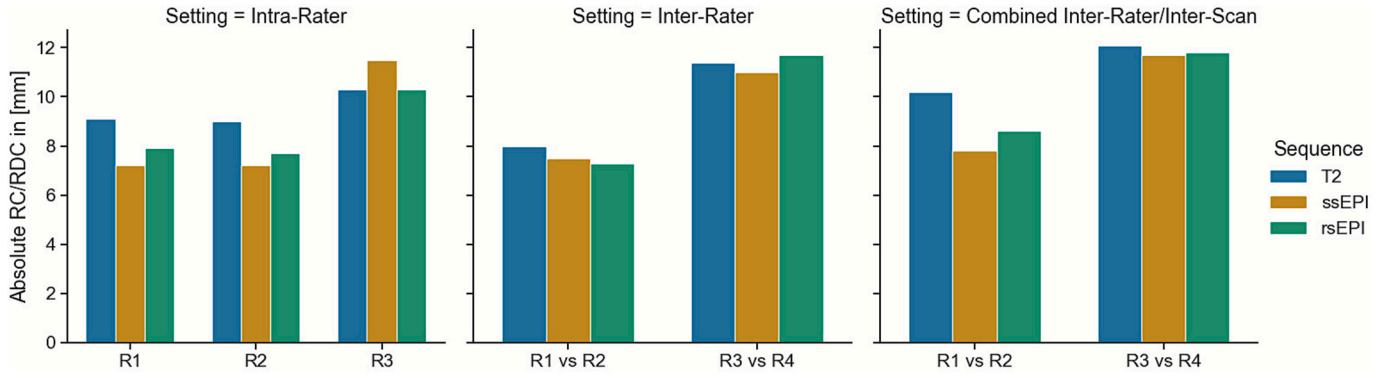

Fig. 4. Absolute repeatability / reproducibility coefficients (RCs/RDCs) of maximum diameter measurements of prostate lesions for the intra-rater, inter-rater and combined inter-rater/inter-scan scenarios for three raters and different sequences.
The combined scenario utilizes measurements from R1/4 of the main exam and compares them to the ones from R2/R3 of the additional exam. *p*-values <0.05 were considered significant.
Abbreviations: rsEPI = diffusion-weighted readout-segmented multi-shot echo planar imaging, ssEPI = diffusion-weighted single-shot echo planar imaging, RC = repeatability coefficient, RDC = reproducibility coefficient.

differences in cohort composition with Diaz de Leon et al. and Marin et al. including only histopathologic index lesions from patients with prostatectomy [35,36] and Rosenkrantz et al. utilizing only index lesions from a non-consecutive cohort [37].

For all sequences, inter-scan and inter-sequence variability could be reduced by performing both measurements in the same reading session, which is intuitively plausible. Mental assessment of lesion form, position and size is performed once and the two scans / sequences are measured with a certain mental co-registration. This implies that, when performing consecutive scans, lesion measurement should be performed in the current and historical scan, which is in accordance with the recommendation of the PRECISE criteria [7].

For all scenarios resulting from different reading sessions, T2WI sequences demonstrated slightly higher variability than DWI. Inflammatory changes or high-grade intraepithelial neoplasia can obscure prostate lesions or at least the borders [38], which may impair delineability and measurement accuracy. According to PI-RADS, TZ lesions are to be measured in the dominant T2WI sequence but, recently, the ability of ADC to characterize prostate lesions in the TZ is being investigated [39,40]. It may be an option to perform size measurements of TZ lesions in the ADC map as well. In concordance, our data did not show significantly different repeatability of MLD measurements in both zones for T2WI and DWI.

Inclusion of all lesions mentioned in the radiologist reports (PI-RDS 2–5) was performed. As only five PI-RADS 2 lesions were included, statistic bias is not expected due to these lesions. Inclusion of potentially less conspicuous PI-RADS 3 lesions in the analysis was essential, as these are the ambiguous lesions, for which potential follow-up imaging may be an alternative to biopsy and where disease progression needs to be identified [41,42]. PI-RADS 4–5 lesions may demonstrate lower measurement variability due to better delineability, however, including only higher category lesions might impact validity of calculated RCs/RDCs for less conspicuous lesions. This reasoning is supported by our findings: PI-RADS≤3 lesions did show higher measurement variability compared to PI-RADS≥4 lesions, however, differences were not statistically significant. Going from lesion level to patient level, i.e., analyzing index lesions only, pooled inter-scan RCs were lower by ~1 mm, suggesting an advantage of a focus on index lesion analysis. However, these findings should be interpreted carefully, as this halved the lesions available for statistical analysis, reducing statistical power. In a (prospective) clinical use case, it would be unknown whether the investigated lesion would be comparable in conspicuity and unambiguity with index lesions in this study.

This study focused on the maximum lesion diameter in the axial plane as the measurement method, which is mentioned as the minimum requirement in the PI-RADS criteria and considered the most common measurement method in clinical practice of the PRECISE expert panel [3,7]. Alternatives are the bi-axial product (lesion cross-section area) or volumetric methods [7,38]. Determination of maximum diameter is fast, well established and used as surrogate parameter for three-dimensional tumor growth and determination of progression [20]. Investigation of other measurement methods was outside the scope of this study and comparison of RCs/RDCs of different methods is limited due to varying dimensional units, e.g. millimeter versus cubic millimeter. The use of (semi-)automatic lesion assessment techniques, possibly supported by or based on artificial intelligence, may improve measurement accuracy due to elimination of aspects like intra-rater variability, and remains to be investigated in the future.

This study is not without limitations. First, a relative threshold in percentage terms would be desirable for lesion change assessment. However, a relative cut-off depends on the lesions size under observation. Instead, the absolute thresholds provided by this study are valid independent from size as the B&A plots demonstrate no funnel-shape [43], indicating identical RC/RDC for different lesion sizes. Second, the number of readers was limited, including two residents, but the congruency and numerical proximity of the calculated metrics from residents and board-certified radiologists (>10 years of experience in prostate MRI) suggest valid measurement results for both groups. Third, sample size of 42 patients is limited but, to our knowledge, the study comprises the largest test-retest cohort of prostate MRI including T2WI and DWI sequences [28,30,32,34,44]. Fourth, it is a retrospective, single scanner study. Prospective multi-center, test-retest studies are necessary to validate our findings in other institutions. Fifth, currently, the PRECISE guidelines suggest size progression at a volume change >50 %. A significant change of MLD according to this study's cut-offs may still result in a lower volume difference. Future studies are necessary to investigate whether smaller changes in MLD measurements could be chosen for a more sensitive adjustment in patient management.

5. Conclusion

To reduce measurement variability during consecutive follow-up scans, prostate lesions should be re-measured on the previous examination. Even with repeat measurement, absolute inter-scan RCs can reach up to 4.8 mm and 5.8 mm for T2WI and DWI, respectively. Diameter changes below these absolute millimeter thresholds may be a result of positioning- and measurement-based variability alone. Our limited data does not demonstrate different measurement variability of sequences depending on the prostate zone. Further studies are necessary to elucidate whether our results also translate to other institutions.

Table 3
Systematic Bias, Coefficient of Variation and Repeatability/Reproducibility Coefficients of Maximum Diameter Measurements of prostate lesions for different rater scenarios: Intra-Rater and Inter-Rater Setting.

	Rater	Sequence	Number of Lesions	Bias (LoAs) [mm]	p adj. for significant bias	Mean [mm]	Relative Bias to Mean [%]	CoV [%]	RC / RDC (95 % CI) [mm]	Relative RC / RDC to Mean [%]
Intra-Rater	R1	T2	82	−0.4 (−7.1, 6.4)	1.0	12.5	−2.9	26.2	9.1 (7.1, 10.6)	72
		ssEPI	85	−0.9 (−6.0, 4.2)	0.26	12.8	−7.1	20.5	7.2 (5.8, 8.2)	56
		rsEPI	71	−0.4 (−6.3, 5.6)	1.0	12.6	−2.8	22.8	7.9 (6.1, 9.4)	63
	R2	T2	82	1.7 (−4.9, 8.3)	<0.01*	10.8	15.3	30.0	9.0 (7.1, 10.5)	83
		ssEPI	85	1.4 (−4.6, 7.4)	0.89	10.8	12.7	24.5	7.2 (5.8, 8.7)	67
		rsEPI	71	1.3 (−4–2, 6.8)	<0.01*	10.9	12.0	25.7	7.7 (6.0, 8.9)	70
	R3	T2	82	−0.6 (−7.9, 6.7)	1.0	13.4	−4.2	27.8	10.3 (8.1, 11.8)	76
		ssEPI	85	−0.8 (−8.9, 7.4)	1.0	13.6	−5.6	30.7	11.5 (9.7, 13.3)	85
		rsEPI	71	−0.4 (−7.7, 7.0)	1.0	13.9	−2.5	27.0	10.3 (8.3, 12.0)	74
Inter-Rater	R1 vs R2	T2	82	0.6 (−5.5, 6.7)	1.0	12.0	5.1	24.1	8.0 (6.3, 9.5)	67
		ssEPI	85	0.9 (−4.6, 6.4)	0.42	11.9	7.9	23.0	7.5 (6.0, 8.7)	63
		rsEPI	71	0.9 (−4.6, 6.4)	1.0	12.0	7.4	22.3	7.3 (5.6, 8.6)	61
	R3 vs R4	T2	82	0.3 (−7.8, 8.4)	1.0	12.9	2.6	32.0	11.4 (9.0, 13.1)	88
		ssEPI	85	0.3 (−7.5, 8.1)	1.0	13.1	2.6	30.6	11.0 (9.1, 12.6)	84
		rsEPI	71	0.7 (−7.6, 9.0)	1.0	13.3	5.4	32.1	11.7 (9.6, 13.8)	88

Linear mixed models were used to test for significance of bias. * p-values <0.05 were considered significant with Holm's method used to correct for multiple comparisons. Note: Lower CoV, RC/RDC and relative RC/RDC denote better repeatability/reproducibility.
Abbreviations: CI = confidence interval, CoV = coefficient of variation, LoAs = limits of agreement, rsEPI = readout-segmented multi-shot echo planar imaging, ssEPI = single-shot echo planar imaging, R1/2/3/4 = Reader 1/2/3/4, RC = repeatability coefficient, RDC = reproducibility coefficient.

CRediT authorship contribution statement

Kevin Sun Zhang: Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Philip Alexander Glemser:** Writing – review & editing, Investigation, Data curation. **Christian Jan Oliver Neelsen:** Writing – review & editing, Investigation, Data curation. **Markus Wennmann:** Writing – review & editing, Conceptualization. **Lukas Thomas Rotkopf:** Writing – review & editing, Methodology. **Nils Netzer:** Writing – review & editing, Data curation. **Clara Meinzer:** Writing – review & editing, Data curation. **Thomas Hielscher:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Vivienn Weru:** Writing – review & editing, Visualization, Formal analysis. **Magdalena Görtz:** Writing – review & editing, Conceptualization. **Albrecht Stenzinger:** Writing – review & editing, Supervision, Resources, Data curation. **Markus Hohenfellner:** Writing – review & editing, Supervision, Resources. **Heinz-Peter Schlemmer:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **David Bonekamp:** Supervision, Resources, Project administration, Data curation, Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing.

Declaration of competing interest

Clara Meinzer reports financial support was provided by Bundesministerium für Wirtschaft und Klimaschutz (BMWK): 01MT21004B. **Magdalena Görtz** reports a relationship with Bayer Vital GmbH that includes: consulting or advisory and speaking and lecture fees. **Magdalena Görtz** reports a relationship with AstraZeneca that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with AstraZeneca that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Novartis that includes: board membership, consulting or advisory,

Table 4
Systematic Bias, Coefficient of Variation and Repeatability/Reproducibility Coefficients of Maximum Diameter Measurements of prostate lesions for the Combined Inter-Rater & Inter-Scan, and Inter-Sequence Setting.

	Rater	Sequence	Number of Lesions	Bias (LoAs) [mm]	p adj. for significant bias	Mean [mm]	Relative Bias to Mean [%]	CoV [%]	RC / RDC (95 % CI) [mm]	Relative RC / RDC to Mean [%]
Combined Inter-Rater and Inter-Scan	R1 vs R2	T2	82	1.2 (−6.6, 9.0)	1.0	12.1	10.0	30.5	10.2 (8.0, 12.1)	84
		ssEPI	85	2.1 (−3.8, 8.0)	<0.01*	12.2	17.3	23.2	7.8 (6.2, 9.1)	64
		rsEPI	52	1.5 (−5.6, 8.6)	0.66	12.6	12.0	27.8	9.6 (7.0, 11.7)	77
	R3 vs R4	T2	82	0.7 (−7.8, 9.3)	1.0	13.2	5.6	33.3	12.1 (9.7, 13.9)	92
		ssEPI	85	0.9 (−7.3, 9.2)	1.0	13.4	6.9	31.8	11.7 (9.7, 13.4)	87
		rsEPI	52	1.4 (−6.9, 9.7)	<0.01*	14.3	9.8	30.0	11.8 (9.3, 14.2)	83
Inter-Sequence (separate reading session)	R1	ssEPI vs rsEPI	71	−0.6 (−6.7, 5.5)	1.0	12.5	−4.9	25.0	8.6 (6.9, 10.0)	69
	R2		71	1.5 (−5.0, 8.0)	0.26	11.0	13.5	24.9	7.4 (5.7, 9.2)	68
	R3		71	−0.6 (−7.9, 6.6)	1.0	13.8	−4.7	26.9	10.2 (7.9, 11.7)	74
Inter-Sequence (same reading session)	R1	ssEPI vs rsEPI	71	−0.2 (−4.4, 4.0)	1.0	12.4	1.6	17	5.7 (4.4, 6.6)	46
	R2		71	0.1 (−4.1, 4.3)	1.0	11.7	0.7	19	5.9 (4.7, 6.9	51
	R3		71	−0.29 (−4.6, 4.0)	1.0	13.6	−2.1	14	5.4 (4.1, 6.5)	39

Linear mixed models were used to test for significance of bias. * p-values <0.05 were considered significant with Holm's method used to correct for multiple comparisons. Note: Lower CoV, RC/RDC and relative RC/RDC denote better repeatability/reproducibility.
Abbreviations: CI = confidence interval, CoV = coefficient of variation, LoAs = limits of agreement, rsEPI = readout-segmented multi-shot echo planar imaging, ssEPI = single-shot echo planar imaging, R1/2/3/4 = Reader 1/2/3/4, RC = repeatability coefficient, RDC = reproducibility coefficient.

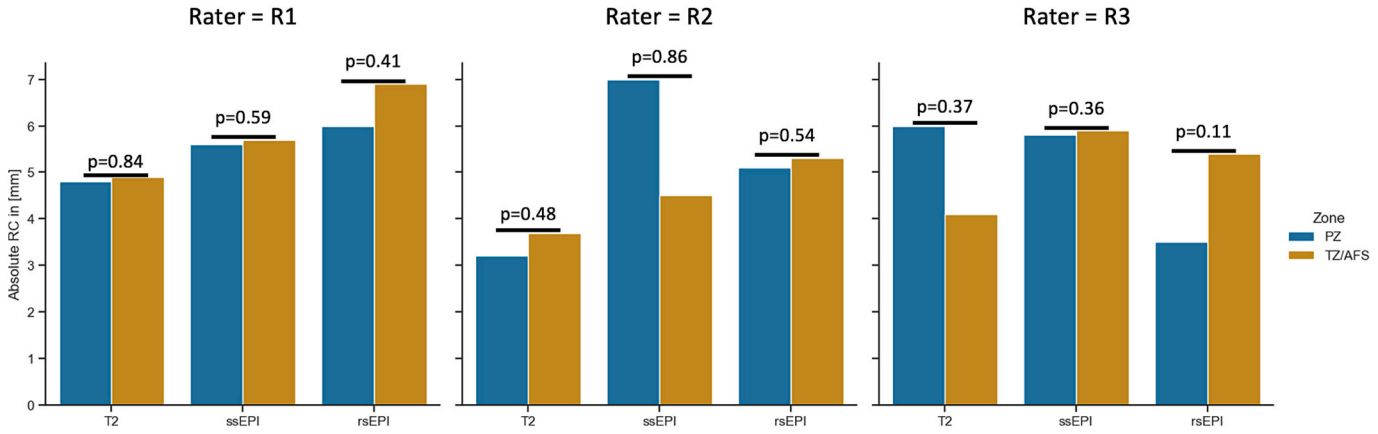

Fig. 5. Absolute repeatability coefficients (RCs) of maximum diameter measurements of prostate lesions for the inter-scan scenario comparing measurements generated in the same reading session and stratified by prostate zones.
Levene's test was used to compare the (inter-scan) variances from the lesions of the PZ and TZ/AFS. * *p*-values <0.05 were considered significant.
Abbreviations: AFS = anterior fibromuscular stroma, PZ = Peripheral Zone, rsEPI = diffusion-weighted readout-segmented multi-shot echo planar imaging, ssEPI = diffusion-weighted single-shot echo planar imaging, RC = repeatability coefficient, TZ = Transition Zone.

Table 5
Pooled Repeatability/Reproducibility Coefficients of Maximum Diameter Measurements of prostate lesions for Reader 1–3 stratified by PI-RADS category (PI-RADS≤3 and PI-RADS≥4).

Setting	Sequence	PI-RADS≤3 RC / RDC (95 % CI) [mm]	PI-RADS≥4 RC / RDC (95 % CI) [mm]	*P* value adjusted according to Levene's Test
Inter-Scan Same-Session	T2	4.9 (3.7, 5.9)	4.6 (3.5, 5.4)	1.0
	ssEPI	5.9 (4.5, 7.2)	5.6 (4.2, 6.9)	1.0
	rsEPI	5.6 (4.0, 7.1)	6.0 (4.4, 8.0)	1.0
Inter-Scan Different Session	T2	10.2 (8.1, 12.3)	8.8 (6.0, 11.6)	1.0
	ssEPI	10.7 (8.5, 13.2)	8.3 (6.5, 10.1)	1.0
	rsEPI	10.3 (7.6, 12.7)	8.6 (5.8, 12.4)	1.0
Intra-Rater	T2	10.4 (8.1, 12.6)	8.3 (5.7, 10.5)	1.0
	ssEPI	10.6 (8.2, 13.3)	7.6 (5.8, 9.5)	0.37
	rsEPI	9.8 (7.5, 12.6)	7.4 (5.3, 9.7)	1.0
Inter-Sequence Same-Session	ssEPI vs. rsEPI	5.6 (4.1, 6.7)	5.8 (4.2, 7.1)	1.0
Inter-Sequence Different-Session	ssEPI vs rsEPI	9.6 (7.0, 12.1)	8.6 (6.2, 11.4)	1.0

Levene's test was used to compare the variances of the two groups (PI-RADS≤3 and PI-RADS≥4). * p-values <0.05 were considered significant with Holm's method used to correct for multiple comparisons.
Abbreviations: CI = Confidence Interval, PI-RADS = Prostate Imaging Reporting and Data System, rsEPI = readout-segmented multi-shot echo planar imaging, ssEPI = single-shot echo planar imaging, RC = repeatability coefficient, RDC = reproducibility coefficient.

speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Illumina Inc that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with Bristol Myers Squibb that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Albrecht Stenzinger** reports a relationship with Roche that includes: consulting or advisory and speaking and lecture fees. **Albrecht Stenzinger** reports a relationship with Thermo Fischer that includes: board membership, consulting or advisory, speaking and lecture fees, and travel reimbursement. **Heinz-Peter Schlemmer** reports a relationship with Bracco Imaging Deutschland GmbH that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. **Heinz-Peter Schlemmer** reports a relationship with Siemens that includes: speaking and lecture fees. **Heinz-Peter Schlemmer** reports a relationship with Bayer Vital GmbH that includes: speaking and lecture fees and travel reimbursement. **David Bonekamp** reports a relationship with Bayer Vital GmbH that includes: speaking and lecture fees. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mri.2025.110578.

References

[1] Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, et al. ESUR prostate MR guidelines 2012. Eur Radiol 2012;22(4):746–57. https://doi.org/10.1007/s00330-011-2377-y.
[2] Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, et al. PI-RADS prostate imaging - reporting and data system: 2015, version 2. Eur Urol 2016;69(1):16–40. https://doi.org/10.1016/j.eururo.2015.08.052.
[3] Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur Urol 2019;76(3):340–51. https://doi.org/10.1016/j.eururo.2019.02.033.
[4] Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate Cancer-2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2021;79(2):243–62. https://doi.org/10.1016/j.eururo.2020.09.042.
[5] Eastham JA, Auffenberg GB, Barocas DA, Chou R, Crispino T, Davis JW, et al. Clinically localized prostate Cancer: AUA/ASTRO guideline, part II: principles of active surveillance, principles of surgery, and follow-up. J Urol 2022;208(1):19–25. https://doi.org/10.1097/ju.0000000000002758.
[6] Moore CM, Giganti F, Albertsen P, Allen C, Bangma C, Briganti A, et al. Reporting magnetic resonance imaging in men on active surveillance for prostate Cancer: the PRECISE recommendations-a report of a European School of Oncology task force. Eur Urol 2017;71(4):648–55. https://doi.org/10.1016/j.eururo.2016.06.011.
[7] Englman C, Maffei D, Allen C, Kirkham A, Albertsen P, Kasivisvanathan V, et al. PRECISE version 2: updated recommendations for reporting prostate magnetic resonance imaging in patients on active surveillance for prostate Cancer. Eur Urol 2024. https://doi.org/10.1016/j.eururo.2024.03.014.
[8] Caglic I, Sushentsev N, Gnanapragasam VJ, Sala E, Shaida N, Koo BC, et al. MRI-derived PRECISE scores for predicting pathologically-confirmed radiological progression in prostate cancer patients on active surveillance. Eur Radiol 2021;31(5):2696–705. https://doi.org/10.1007/s00330-020-07336-0.
[9] Giganti F, Pecoraro M, Stavrinides V, Stabile A, Cipollari S, Sciarra A, et al. Interobserver reproducibility of the PRECISE scoring system for prostate MRI on active surveillance: results from a two-Centre pilot study. Eur Radiol 2020;30(4):2082–90. https://doi.org/10.1007/s00330-019-06557-2.
[10] Giganti F, Stabile A, Stavrinides V, Osinibi E, Retter A, Orczyk C, et al. Natural history of prostate cancer on active surveillance: stratification by MRI using the PRECISE recommendations in a UK cohort. Eur Radiol 2021;31(3):1644–55. https://doi.org/10.1007/s00330-020-07256-z.
[11] Ullrich T, Arsov C, Quentin M, Mones F, Westphalen AC, Mally D, et al. Multiparametric magnetic resonance imaging can exclude prostate cancer progression in patients on active surveillance: a retrospective cohort study. Eur Radiol 2020;30(11):6042–51. https://doi.org/10.1007/s00330-020-06997-1.
[12] Zhang KS, Neelsen CJO, Wennmann M, Glemser PA, Hielscher T, Weru V, et al. Same-day repeatability and between-sequence reproducibility of mean ADC in PI-RADS lesions. Eur J Radiol 2023;165:110898. https://doi.org/10.1016/j.ejrad.2023.110898.
[13] Zhang KS, Neelsen CJO, Wennmann M, Hielscher T, Kovacs B, Glemser PA, et al. In vivo variability of MRI radiomics features in prostate lesions assessed by a test-retest study with repositioning. Sci Rep 2025;15(1):29703. https://doi.org/10.1038/s41598-025-09989-7.
[14] Parker RA, Weir CJ, Rubio N, Rabinovich R, Pinnock H, Hanley J, et al. Application of mixed effects limits of agreement in the presence of multiple sources of variability: exemplar from the comparison of several devices to measure respiratory rate in COPD patients. PloS One 2016;11(12):e0168321. https://doi.org/10.1371/journal.pone.0168321.
[15] Zou GY. Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. Stat Methods Med Res 2013;22(6):630–42. https://doi.org/10.1177/0962280211402548.
[16] Bland JM, Altman DG. Measurement error. BMJ 1996;312(7047):1654. https://doi.org/10.1136/bmj.312.7047.1654.
[17] Bradley EL, Blackwood LG. Comparing paired data - a simultaneous test for means and variances. Am Stat 1989;43(4):234–5. https://doi.org/10.2307/2685368.
[18] Gregg M, Creuziger A, Datta S, Lorenz D. Two tests of variance homogeneity for clustered data where group size is informative. J Stat Comput Simul 2025;95(3):490–506. https://doi.org/10.1080/00949655.2024.2430692.
[19] Holm S. A simple sequentially Rejective multiple test procedure. Scand J Stat 1979;6(2):65–70.
[20] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45(2):228–47. https://doi.org/10.1016/j.ejca.2008.10.026.
[21] Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol 2010;28(11):1963–72. https://doi.org/10.1200/jco.2009.26.3541.
[22] Cheson BD, Ansell S, Schwartz L, Gordon LI, Advani R, Jacene HA, et al. Refinement of the Lugano classification lymphoma response criteria in the era of

immunomodulatory therapy. Blood 2016;128(21):2489–96. https://doi.org/10.1182/blood-2016-05-718528.
[23] Harder FN, Heming CAM, Haider MA. mpMRI interpretation in active surveillance for prostate Cancer-an overview of the PRECISE score. Abdom Radiol (NY) 2023;48(7):2449–55. https://doi.org/10.1007/s00261-023-03912-2.
[24] Wennmann M, Grözinger M, Weru V, Hielscher T, Rotkopf LT, Bauer F, et al. Test-retest, inter- and intra-rater reproducibility of size measurements of focal bone marrow lesions in MRI in patients with multiple myeloma. Br J Radiol 2023;96(1145):20220745. https://doi.org/10.1259/bjr.20220745.
[25] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology 2009;252(1):263–72. https://doi.org/10.1148/radiol.2522081593.
[26] Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol 2011;29(23):3114–9. https://doi.org/10.1200/JCO.2010.33.7071.
[27] McErlean A, Panicek DM, Zabor EC, Moskowitz CS, Bitar R, Motzer RJ, et al. Intra- and interobserver variability in CT measurements in oncology. Radiology 2013;269(2):451–9. https://doi.org/10.1148/radiology.13122665.
[28] Fedorov A, Vangel MG, Tempany CM, Fennessy FM. Multiparametric magnetic resonance imaging of the prostate: repeatability of volume and apparent diffusion coefficient quantification. Invest Radiol 2017;52(9):538–46. https://doi.org/10.1097/RLI.0000000000000382.
[29] Michoux NF, Ceranka JW, Vandemeulebroucke J, Peeters F, Lu P, Absil J, et al. Repeatability and reproducibility of ADC measurements: a prospective multicenter whole-body-MRI study. Eur Radiol 2021;31(7):4514–27. https://doi.org/10.1007/s00330-020-07522-0.
[30] Boss MA, Snyder BS, Kim E, Flamini D, Englander S, Sundaram KM, et al. Repeatability and reproducibility assessment of the apparent diffusion coefficient in the prostate: a trial of the ECOG-ACRIN research group (ACRIN 6701). J Magn Reson Imaging 2022;56(3):668–79. https://doi.org/10.1002/jmri.28093.
[31] Sadinski M, Medved M, Karademir I, Wang S, Peng Y, Jiang Y, et al. Short-term reproducibility of apparent diffusion coefficient estimated from diffusion-weighted MRI of the prostate. Abdom Imaging 2015;40(7):2523–8. https://doi.org/10.1007/s00261-015-0396-x.
[32] Rogers HJ, Singh S, Barnes A, Obuchowski NA, Margolis DJ, Malyarenko DI, et al. Test-retest repeatability of ADC in prostate using the multi b-value VERDICT acquisition. Eur J Radiol 2023;162:110782. https://doi.org/10.1016/j.ejrad.2023.110782.
[33] Lo WC, Bittencourt LK, Panda A, Jiang Y, Tokuda J, Seethamraju R, et al. Multicenter repeatability and reproducibility of MR fingerprinting in phantoms and in prostatic tissue. Magn Reson Med 2022;88(4):1818–27. https://doi.org/10.1002/mrm.29264.
[34] Hiremath A, Shiradkar R, Merisaari H, Prasanna P, Ettala O, Taimen P, et al. Test-retest repeatability of a deep learning architecture in detecting and segmenting clinically significant prostate cancer on apparent diffusion coefficient (ADC) maps. Eur Radiol 2021;31(1):379–91. https://doi.org/10.1007/s00330-020-07065-4.
[35] Marin L, Ezziane M, Comperat E, Mozer P, Cancel-Tassin G, Coté JF, et al. Comparison of semi-automated and manual methods to measure the volume of prostate cancer on magnetic resonance imaging. Diagn Interv Imaging 2017;98(5):423–8. https://doi.org/10.1016/j.diii.2017.02.004.
[36] Diaz de Leon A, Leyendecker JR, Otero-Muinelo S, Grewal H, Xi Y, Francis F, et al. Reproducibility of index lesion size and mean apparent diffusion coefficient values measured by prostate multiparametric MRI: correlation with whole-mount sectioning of specimens. AJR Am J Roentgenol 2018;211(4):783–8. https://doi.org/10.2214/ajr.17.19172.
[37] Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, et al. Interobserver reproducibility of the PI-RADS version 2 lexicon: a multicenter study of six experienced prostate radiologists. Radiology 2016;280(3):793–804. https://doi.org/10.1148/radiol.2016152542.
[38] Giganti F, Stavrinides V, Stabile A, Osinibi E, Orczyk C, Radtke JP, et al. Prostate cancer measurements on serial MRI during active surveillance: it's time to be PRECISE. Br J Radiol 2020;93(1116):20200819. https://doi.org/10.1259/bjr.20200819.
[39] Bonekamp D, Kohl S, Wiesenfarth M, Schelb P, Radtke JP, Gotz M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: comparison to ADC values. Radiology 2018;289(1):128–37. https://doi.org/10.1148/radiol.2018173064.
[40] Panda A, Obmann VC, Lo WC, Margevicius S, Jiang Y, Schluchter M, et al. MR fingerprinting and ADC mapping for characterization of lesions in the transition zone of the prostate gland. Radiology 2019;292(3):685–94. https://doi.org/10.1148/radiol.2019181705.
[41] Boschheidgen M, Schimmöller L, Doerfler S, Al-Monajjed R, Morawitz J, Ziayee F, et al. Single center analysis of an advisable control interval for follow-up of patients with PI-RADS category 3 in multiparametric MRI of the prostate. Sci Rep 2022;12(1):6746. https://doi.org/10.1038/s41598-022-10859-9.
[42] Schoots IG. MRI in early prostate cancer detection: how to manage indeterminate or equivocal PI-RADS 3 lesions? Transl Androl Urol 2018;7(1):70–82. https://doi.org/10.21037/tau.2017.12.31.
[43] Kopp-Schneider A, Hielscher T. How to evaluate agreement between quantitative measurements. Radiother Oncol 2019;141:321–6. https://doi.org/10.1016/j.radonc.2019.09.004.
[44] McGarry SD, Brehler M, Bukowy JD, Lowman AK, Bobholz SA, Duenweg SR, et al. Multi-site concordance of diffusion-weighted imaging quantification for assessing prostate Cancer aggressiveness. J Magn Reson Imaging 2022;55(6):1745–58. https://doi.org/10.1002/jmri.27983.