scientific reports

Check for updates

OPEN

Multicenter analysis for inter reader agreement of PIQUAL score v2 for basic readers in prostate MRI

Mujgan Orman[1], Sabri Sirolu[2], Mustafa Ege Seker[3], Deniz Alis[4]✉, Mustafa Said Kartal[5], Yeliz Basar[1], Nurper Denizoglu[6], Serpil Kurtcan[1], Aydan Arslan[7], Ilkay Oksuz[8] & Ercan Karaarslan[4]

The Prostate Imaging Quality (PI-QUAL) V2 is a simplified update of PI-QUAL V1 for assessing image quality in multiparametric prostate MRI (mpMRI), aimed at improving clinical use and reproducibility. This study evaluated inter-reader agreement of PI-QUAL V2 among basic prostate readers using multicenter mpMRI data and compared it to PI-QUAL V1. Five readers assessed T2-weighted, DWI (including ADC maps), and DCE images from 274 men (median age: 71 years, IQR: 60–78) across five centers. Inter-reader agreement was measured using weighted Gwet's AC1 coefficients and compared using bootstrapped confidence intervals. For PI-QUAL V1, coefficients ranged from 0.659 to 0.855, indicating substantial-to-almost perfect agreement. For PI-QUAL V2, coefficients ranged from 0.821 to 0.897, showing almost perfect agreement. All 10 pairwise comparisons achieved almost perfect agreement for PI-QUAL V2, compared to only 3 of 10 for V1. Additionally, three comparisons showed non-overlapping confidence intervals, indicating statistically significant improvements with V2. PI-QUAL V2 demonstrated consistently higher inter-reader agreement than V1, with basic prostate readers achieving almost perfect agreement across multicenter mpMRI data. These findings highlight the improved reliability and clinical applicability of PI-QUAL V2.

Keywords Prostate, MRI, Prostate Cancer, PI-QUAL, Image quality

Multiparametric magnetic resonance imaging (mpMRI) of the prostate has emerged as the preferred diagnostic tool for biopsy-naïve patients suspected of having clinically significant prostate cancer (csPCa), outperforming traditional screening methods like prostate-specific antigen (PSA) testing, which can lead to overdiagnosis and overtreatment[1–3]. However, concerns have arisen regarding the variability and often suboptimal image quality of prostate mpMRI scans, which is crucial for the accurate detection of csPCa[4,5]. While the Prostate Imaging Reporting and Data System (PI-RADS) guidelines set minimum technical requirements for prostate MRI acquisition, adherence to these guidelines does not always ensure high-quality scans[6–8].

To address the need for a standardized evaluation of prostate MRI image quality, Giganti et al. introduced the Prostate Imaging Quality (PI-QUAL) V1 in 2020, using data from the PRECISION trial[9,10]. PI-QUAL V1 combines an objective assessment of technical parameters, such as field-of-view and slice thickness, with a subjective evaluation based on human reader perceptions. Studies have demonstrated that image quality assessed via PI-QUAL V1 directly influences cancer detection and biopsy planning. Furthermore, PI-QUAL V1 has shown moderate-to-good inter-reader agreement among both expert and basic prostate readers[11,12].

Despite these advantages, PI-QUAL V1 has several limitations[13,14]. A significant limitation is that it is suitable only for mpMRI examinations that include dynamic contrast-enhanced (DCE) sequences, thus excluding examinations without intravenous contrast medium, such as bi-parametric MRI (bpMRI). Furthermore, PI-QUAL V1 does not enable an objective evaluation of image quality distinct from biopsy implications, restricting

[1]Department of Radiology, Acibadem Healthcare Group, Istanbul 34457, Turkey. [2]Department of Radiology, Istanbul Sisli Hamidiye Etfal Training and Research Hospital, Istanbul 34396, Turkey. [3]School of Medicine, Department of Radiology, University of Wisconsin-Madison, Madison, WI 53705, USA. [4]School of Medicine, Department of Radiology, Acibadem Mehmet Ali Aydinlar University, Istanbul 34457, Turkey. [5]School of Medicine, Cumhuriyet University, Sivas 581407, Turkey. [6]Department of Radiology, Sultan 2. Abdulhamid Han Training and Research Hospital, Istanbul 34668, Turkey. [7]Department of Radiology, Umraniye Training and Research Hospital, Istanbul 34764, Turkey. [8]Department of Computer Engineering, Istanbul Technical University, Istanbul 34467, Turkey. ✉email: drdenizalis@gmail.com

its use in other clinical scenarios such as tumor staging, active surveillance, population screening, and follow-up of patients with prior negative or positive scans.

To address these limitations, the European Society of Urogenital Radiology (ESUR) prostate cancer working group recently developed an updated version of the PI-QUAL score, known as PI-QUAL V2[15]. This revised scoring system aims to improve clinical usability and consistency by (i) including MRI examinations without intravenous contrast, (ii) streamlining the scoring process, and (iii) ensuring reproducible assessments of image quality for a broader range of patients in routine clinical practice. The introduction of PI-QUAL V2 seeks to further standardize and enhance the reliability of prostate MRI quality evaluations.

In this study, we assessed the inter-reader agreement of the PI-QUAL V2 scoring system among five basic prostate readers from different centers on a multicenter dataset. Furthermore, we compared the performance of PI-QUAL V2 with PI-QUAL V1 score to determine whether the simplified version improves the inter-rater agreement among basic prostate readers.

Methods

Study sample

This retrospective study received approval from the institutional review boards of all participating centers, with a waiver of informed consent due to the use of anonymized medical data. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. We retrospectively reviewed records from seven hospitals to identify consecutive patients who underwent prostate MRI examinations between January 2023 and December 2023. These patients were either suspected of having csPCa—as indicated by elevated prostate-specific antigen (PSA) levels or abnormal findings on digital rectal examination—or were undergoing active surveillance.

We excluded patients who had a prior history of treatment for csPCa ($n = 23$) and those who underwent bi-parametric MRI scans without intravenous contrast administration ($n = 167$).

Multiparametric MRI protocols

The mpMRI exams were performed using either 1.5 Tesla (Essenza or Avanto-fit, Siemens Healthcare, Erlangen, Germany) or 3 Tesla (Prisma-fit, Vida, Siemens Healthcare, Erlangen, Germany) MRI systems. All exams utilized phased-array surface coils and included the following sequences following PI-RADS V2.1.: bi-planar T2-weighted images (T2 WI) acquired in axial, sagittal, or coronal planes; diffusion-weighted imaging (DWI), including ADC maps and high b-value images (either directly acquired or calculated with b-values of ≥ 1,400 s/mm²); and dynamic contrast-enhanced (DCE) imaging. Detailed information about the MRI protocols used at each center is provided in Supplementary Document 1.

PI-QUAL V1 and V2 assessment

Five radiologists (S.S., Y.B., N.D., S.K., A.A.) from five separate medical centers participated in the quality assessment of prostate mpMRI scans using the PI-QUAL V1 and V2 scoring system. The evaluations were conducted on a specialized browser-based platform (https://matrix.md.ai) equipped with a 6-megapixel diagnostic color monitor (Radiforce RX 660, EIZO). To maintain consistency, all images were reviewed in DICOM format.

At the beginning of the study, each reader had experience interpreting more than 400 but fewer than 1,000 prostate MRI cases and was reading up to 300 mpMRI scans annually. According to established guidelines[4], all participants qualified as "basic prostate readers."

To ensure an unbiased evaluation, the readers were blinded to all clinical information during the assessment. Prior to independent scoring, all five readers completed a structured curriculum comprising three two-hour online workshops and one three-hour in-person session (total contact time ≈ 9 h), led by a radiologist with over 20 years of experience in prostate imaging (E.K.). The training dataset included 30 mpMRI exams deliberately balanced across image quality strata: 10 inadequate (PI-QUAL 1–2), 10 acceptable (PI-QUAL 3), and 10 optimal (PI-QUAL 4–5). All cases were sourced from participating centers but excluded from the final study dataset.

During each session, the expert first reviewed relevant anatomical landmarks and PI-QUAL criteria, then facilitated an open discussion as basic readers independently scored six cases in rotating order. Importantly, no reference scores were disclosed until after consensus was recorded to mitigate anchoring bias. No additional feedback or calibration was provided after the training phase concluded. The PI-QUAL V1 and V2 score sheet templates used in the study is presented in Figs. 1 and 2.

The readers assessed the images for both the PI-QUAL V1 and V2 scoring systems during the same sessions over a period of two months. They were asked to complete the cases in twelve separate sessions, with approximately 25 cases per session.

Statistical analyses

All statistical analyses were performed using the SciPy library in Python V3. Continuous variables are presented as medians with interquartile ranges, while categorical and ordinal variables are expressed as frequencies and percentages. The PI-QUAL V1 and V2 scores assigned by the readers were calculated and compared on a per-scan basis. To evaluate inter-reader agreement among the readers for PI-QUAL scoring, Gwet's AC1 statistics were used instead of kappa due to its robustness against prevalence effects and sensitivity to imbalanced category distributions[11,12,16].

For PI-QUAL V1, Gwet's AC1 was calculated with linear and full weights after harmonizing PI-QUAL V1 scores to a 3-point scale. The harmonization process mapped scores 1–2 to V2 category 1 (inadequate), score 3 to V2 category 2 (acceptable), and scores 4–5 to V2 category 3 (optimal), preserving clinical meaning while

Scan & site number:

PRECISION

Prostate Imaging QUALity control (PI-QUAL) scoring sheet

PI-QUAL score	Criteria	Clinical implications
1	All mpMRI sequences are below the minimum standard of diagnostic quality	It is NOT possible to rule in all significant lesions [§] It is NOT possible to rule out all significant lesions [§]
2	Only one mpMRI sequence is of acceptable diagnostic quality	
3	At least two mpMRI sequences taken together are of diagnostic quality	It is possible to rule in all significant lesions It is NOT possible to rule out all significant lesions
4	Two or more mpMRI sequences are independently of diagnostic quality	It is possible to rule in all significant lesions It is possible to rule out all significant lesions
5	All mpMRI sequences are of optimal diagnostic quality	

[§] Therefore reports should not include PI-RADS or Likert scores

Please (✓) if present: (note: '*adequate*' means compliant with the technical specifications reported in PI-RADS v. 2 guidelines) *

T2-WI

Technical parameters	
Axial plane	
Sagittal or coronal plane	
Adequate field of view	
Adequate in-plane resolution	
Adequate slice thickness	
Z-axis correctly positioned	
Visual assessment	
Capsule clearly delineated	
Seminal vesicles clearly delineated	
Ejaculatory ducts clearly delineated	
Neurovascular bundles clearly delineated	
Sphincter muscle clearly delineated	
Absence of artefacts (e.g. movement)	

DWI

Technical parameters	
Axial plane matching T2-WI	
Adequate field of view	
Adequate in-plane resolution	
Adequate slice thickness	
Multiple [> 2] *b* values acquired	
High *b* value (synthesised or acquired)	
Visual assessment	
Adequate ADC map	
Absence of artefacts (e.g. rectal air)	

DCE

Technical parameters	
Axial plane matching T2-WI	
Adequate field of view	
Adequate in-plane resolution	
Adequate slice thickness	
Pre-contrast T1-WI available	
Fat suppression/subtraction	
Adequate temporal resolution [≤ 10 sec]	
Adequate total observation rate [≥ 2min]	
Visual assessment	
Capsular vessels clearly delineated	
Vessels in the Alcock's canal clearly delineated	
Absence of artefacts (e.g. movement)	

Is T2-WI of diagnostic quality? ☐ Yes ☐ No

Is DWI of diagnostic quality? ☐ Yes ☐ No

Is DCE of diagnostic quality? ☐ Yes ☐ No

PI-QUAL score: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Comments:

Date:

Reporting Radiologist:

Signed:

* *Weinreb JC, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2016;69:16–40.*

Fig. 1. Prostate Imaging Quality (PI-QUAL) V1 scoring sheet. Reprinted from Giganti, Francesco, et al. "Prostate Imaging Quality (PI-QUAL): a new quality control scoring system for multiparametric magnetic resonance imaging of the prostate from the PRECISION trial." European urology oncology 3.5 (2020): 615–619. Copyright 2020 Elsevier.

enabling direct comparisons between the two systems. For PI-QUAL V2, Gwet's AC1 was calculated using quadratic weights to account for the ordinal nature of the scoring system.

The Gwet's AC1 values were interpreted according to the following scale: less than 0 indicates poor agreement; 0 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, almost perfect agreement. Confidence intervals for Gwet's AC1 were derived using bootstrapping with 1,000 iterations, providing robust estimates of variability without reliance on parametric assumptions.

Results

A total of 274 men were enrolled in this study, with a median age of 71 years (interquartile range [IQR], 60–78 years). Of the mpMRI examinations performed, 176 out of 274 were conducted using a 1.5 Tesla (T) magnet, while the remaining 98 were obtained using a 3 T magnet. The distribution of scans among the centers was as follows: Center 1 contributed 49 exams, Center 2 contributed 20 exams, Center 3 contributed 49 exams, Center 4 contributed 17 exams, Center 5 contributed 44 exams, Center 6 contributed 46 exams, and Center 7 contributed 49 scans.

PI-QUAL V1 and V2 scores

The readers assigned a PI-QUAL V1 score of ≥ 3 (i.e., sufficient to rule in all clinically significant lesions) for 94.9% of the assessments. For PI-QUAL V2, the readers assigned a score of ≥ 2 (i.e., acceptable for interpretation) for 75.1% of the exams. Tables 1 and 2 present the PI-QUAL V1 and V2 scores for each reader.

For PI-QUAL V1, Gwet's AC1 values ranged from 0.659 (95% CI: 0.586–0.732) (R3 vs. R4) to 0.855 (95% CI: 0.806–0.905) (R2 vs. R5), indicating substantial-to-almost perfect agreement. In contrast, PI-QUAL V2 demonstrated overall higher agreement, with AC1 values ranging from 0.821 (95% CI: 0.783–0.861) (R3 vs. R5) to 0.897 (95% CI: 0.866–0.929) (R4 vs. R5). Table 3 shows Gwet's AC1 values for each pair of readers.

For PI-QUAL V1, 7 out of 10 comparisons showed substantial agreement, while the remaining comparisons showed almost perfect agreement. In contrast, all comparisons for PI-QUAL V2 demonstrated almost perfect agreement. Additionally, for each pair of readers, Gwet's AC1 values were higher for PI-QUAL V2 compared with V1. Notably, 3 out of 10 comparisons had non-overlapping bootstrapped 95% confidence intervals between PI-QUAL V1 and V2, indicating statistically significant improvements in inter-reader agreement.

Figure 3 illustrates the pairwise inter-reader agreements among readers for PI-QUAL V1 and V2 scores. Figures 4 and 5 display the PI-QUAL V1 and V2 scores for representative cases. Confusion matrices of pairwise PI-QUAL scores for each reader are provided in Supplementary Document 2.

PI-QUAL v2 scoring sheet

MRI without intravenous contrast medium

T2-WI	DWI	PI-QUAL score	Remarks	General clinical implication
≤ 2	≤ 2	1	-	Inadequate scan: scan should be repeated
3 or 4	3 or 4	2	No: ≤ 2 /4 for T2-WI and DWI	Acceptable scan: consider repeat scan
4	4	3	Full scores for T2-WI and DWI	Optimal scan: scan of optimal diagnostic quality

Multiparametric MRI

T2-WI	DWI	DCE	PI-QUAL score	Remarks	General clinical implication
≤ 2	≤ 2	+	1	-	Inadequate scan: scan should be repeated
3 or 4	3 or 4		2	No: ≤ 2 /4 for T2-WI and DWI	Acceptable scan: consider repeat scan
4	4	-	3	Full scores for T2-WI and DWI	Optimal scan: scan of optimal diagnostic quality

'+' : both criteria for DCE are satisfied and at least one sequence (either T2-WI or DWI) must score 4/4
'-' : either only one criterion or no criteria for DCE are satisfied

Please (✓) if present:

T2-WI

Essential requirement before proceeding (equals 0/4 if not met):
Slice thickness: 3 mm
Axial T2-WI: adequate signal-to-noise ratio (SNR) in all parts of the images
Axial T2-WI: ability to clearly delineate relevant structures in the prostate
Axial T2-WI: absence of significant artefacts in the prostatic region
Sagittal OR coronal: adequate SNR and image resolution AND absence of significant artefacts
Total score for T2-WI / 4

DWI

Essential requirement before proceeding (equals 0/4 if not met):
Slice thickness: ≤ 4 mm
High *b* value sequence (≥ 1,400 s/mm²)
ADC map using at least two *b* values up to 1,000 s/mm²
Adequate contrast and SNR on high *b* value images
Adequate range of contrast to differentiate TZ/BPH from PZ on the ADC maps
Absence of significant artefacts in the prostatic region
Anatomical matching of the ADC map / high *b* value sequence to the axial T2-WI
Total score for DWI / 4

DCE

Essential requirement before proceeding (equals '–' if not met):
Slice thickness: 3 mm
Temporal resolution: ≤ 15 seconds
Fat saturation (or include post-processing, e.g. subtraction / heat maps)
Absence of significant artefacts in the prostatic region and appropriate bolus enhancement
Ability to identify anatomical structures (e.g. capsular vessels or pudendal artery)
Total score for DCE ('+' only when both criteria are met) + / -

PI-QUAL score ☐ 1 ☐ 2 ☐ 3

Fig. 2. Prostate Imaging Quality (PI-QUAL) V2 scoring sheet. Reprinted from de Rooij, Maarten, et al. "PI-QUAL version 2: an update of a standardized scoring system for the assessment of image quality of prostate MRI." European Radiology (2024). Copyright 2024 Springer Nature.

PI-QUAL V1 Score	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Overall Scores
1	0 (0%)	2 (0.7%)	8 (2.9%)	3 (1.1%)	0 (0%)	1%
2	12 (4.4%)	11 (4%)	16 (5.8%)	6 (2.2%)	11 (4%)	4.1%
3	24 (8.8%)	16 (5.8%)	27 (9.9%)	51 (18.6%)	15 (5.5%)	9.7%
4	44 (16%)	47 (17.2%)	48 (17.5%)	4 (1.5%)	74 (27%)	15.8%
5	194 (70.8%)	198 (72.3%)	175 (63.9%)	210 (76.6%)	174 (63.5%)	69.4%

Table 1. Prostate imaging quality (PI-QUAL) V1 scores for each reader.

Discussion

This study evaluated the inter-reader agreement among five basic prostate readers in assigning PI-QUAL V1 and V2 scores across multi-center data, aiming to determine whether the updated and simplified PI-QUAL V2 achieved higher inter-rater agreement. To the best of our knowledge, this research provides initial evidence of

PI-QUAL V2 Score	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Overall Scores
1	71 (25.9%)	61 (22.6%)	84 (30.7%)	60 (21.9%)	66 (24.1%)	24.9%
2	166 (60.6%)	168 (61.3%)	147 (53.6%)	176 (64.2%)	166 (60.6%)	60.2%
3	37 (13.5%)	44 (16.1%)	43 (15.7%)	38 (13.9%)	42 (15.3%)	14.9%

Table 2. Prostate imaging quality (PI-QUAL) V2 scores for each reader.

PI-QUAL Score	PI-QUAL V1 (95% CI)	PI-QUAL V2(95% CI)	Increment
Reader 1vs2	0.847 (0.796–0.897)	0.862 (0.831–0.895)	+ 0.015
Reader 1vs3	0.762 (0.699–0.826)	0.835 (0.798–0.874)	+ 0.073
Reader 1vs4	**0.713 (0.646–0.781)**	**0.839 (0.802–0.877)**	**+ 0.126**
Reader 1vs5	0.815 (0.803–0.901)	0.87 (0.839–0.902)	+ 0.055
Reader 2vs3	0.768 (0.705–0.833)	0.852 (0.816–0.889)	+ 0.084
Reader 2vs4	0.738 (0.674–0.804)	0.835 (0.797–0.874)	+ 0.097
Reader 2vs5	0.855 (0.806–0.905)	0.85 (0.817–0.884)	−0.005
Reader 3vs4	**0.659 (0.586–0.732)**	**0.831 (0.793–0.87)**	**+ 0.172**
Reader 3vs5	0.728 (0.659–0.797)	0.821 (0.783–0.861)	+ 0.093
Reader 4vs5	**0.773 (0.714–0.832)**	**0.897 (0.866–0.929)**	**+ 0.124**

Table 3. The comparison of reader consistency of Gwet's AC1 scores of each reader for prostate imaging quality (PI-QUAL) V1 and V2. *Bold indicates non-overlapping bootstrapped 95% confidence intervals between PI-QUAL V1 and V2, showing a statistically higher difference.

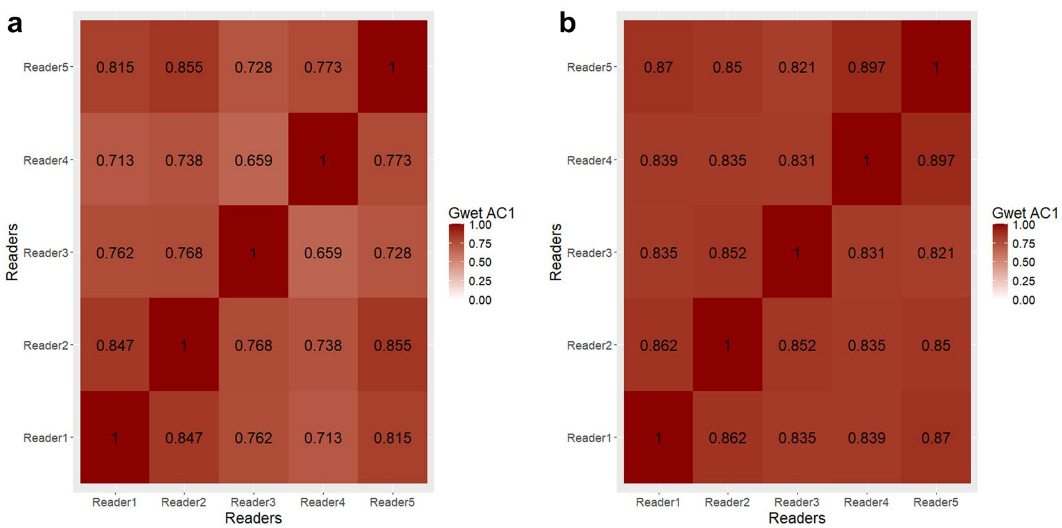

Fig. 3. The pair-wise Gwet's AC_1 scores of readers in assigning Prostate Imaging Quality (PI-QUAL) V1 and V2 scores. (**a**) The Gwet's AC_1 scores of the basic prostate readers range from 0.659 to 0.855, indicating substantial-to-almost perfect levels of inter-reader agreement in assigning PI-QUAL V1 scores. (**b**) The Gwet's AC_1 scores of the readers range from 0.821 to 0.897, indicating almost perfect level of inter-reader agreement in assigning PI-QUAL V2 scores.

inter-rater agreement for PI-QUAL V2 among basic prostate readers using multi-center data and compares the agreement levels between PI-QUAL V1 1 and 2.

We found that the inter-rater agreement among basic prostate readers ranged from substantial-to-almost perfect for PI-QUAL V1 scores and almost perfect for PI-QUAL V2 scores, with overall higher inter-rater agreement for each pair of readers in PI-QUAL V2. Our findings suggest that the simplified three-point scale of PI-QUAL V2, focusing on essential technical and qualitative parameters, maintains strong inter-reader agreement while also being easier to apply in clinical practice. While some scoring discrepancies persisted, PI-QUAL V2 offered a more consistent and reproducible framework, resulting in more aligned image quality assessments among readers with varying levels of experience. We believe that the improved inter-reader agreement observed

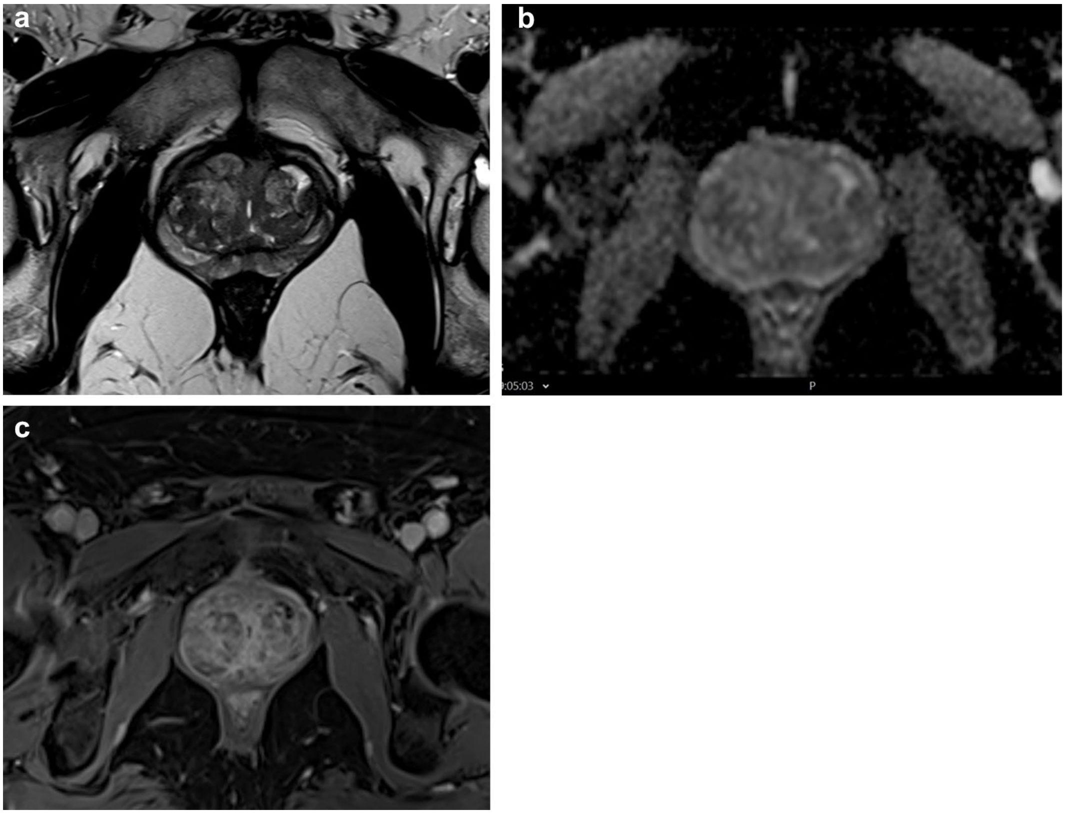

Fig. 4. A 70-year-old man who underwent multi-parametric prostate MRI at 3 T. **(a)** On the axial T2-weighted image, the signal-to-noise ratio is adequate, allowing clear delineation of anatomical structures. No artifacts are present, and the signal-to-noise ratio remains sufficient on both coronal and sagittal T2 W images (not shown). **(b)** The anatomically matched axial apparent diffusion coefficient (ADC) map is artifact-free, with adequate contrast and signal-to-noise ratio, clearly delineating the peripheral and transitional zones. **(c)** The vessels are clearly delineated, with no visible artifacts on the dynamic-contrast-enhanced image. All readers assigned a PI-QUAL score of 5 for PI-QUAL V1 and a score of 3 for PI-QUAL V2 in this case.

with PI-QUAL V2 suggests that it can serve as a more consistent and reliable tool for assessing prostate MRI quality, potentially leading to more standardized care and improved diagnostic outcomes.

Several studies have evaluated inter-reader agreement for PI-QUAL V1, yielding varied results. In the study by Giganti et al., expert readers achieved excellent pairwise inter-reader agreement with a kappa score of 0.82, highlighting PI-QUAL V1's reliability in the hands of experienced readers who can more consistently assess the technical quality of prostate MRI. However, when basic prostate readers assessed PI-QUAL V1, as shown in the study by Basar et al., the kappa scores ranged from 0.656 to 0.786—reflecting good, though lower, levels of agreement compared to expert readers[16]. Additionally, Karanasios et al. observed moderate inter-reader agreement, with kappa scores of 0.47 between a junior and senior reader and 0.52 between senior readers, indicating greater variability in image quality interpretation[12].

In the study by Ponsiglione et al. on PI-QUAL V2, the authors examined both inter-reader and intra-reader agreement among readers of varying experience levels using the PI-QUAL V2 scoring system[17]. The study found that overall inter-reader agreement was moderate, with higher consistency among expert readers than basic and beginner readers. Intra-reader agreement ranged from moderate to perfect, with consistency improving with reader experience. Unlike our study, Ponsiglione et al. did not directly compare PI-QUAL V1 and V2 scores among readers.

Our study employed weighted Gwet's AC1 to assess inter-reader agreement, whereas Ponsiglione et al. used non-weighted Gwet's AC1. Although both studies utilized the same metric, they handled the scoring systems differently. Ponsiglione et al. dichotomized both versions of PI-QUAL, effectively reducing the granularity of the scales. In contrast, our study preserved the original three-level (PI-QUAL V2) and five-level (PI-QUAL V1) structures, applying weight normalization to Gwet's AC1 to account for varying category distributions. This harmonization not only ensured consistency between scales but also allowed meaningful clinical comparisons, aligning the scoring with diagnostic significance. However, this methodological difference limits the direct comparability of findings between the two studies, and caution should be exercised when interpreting agreement levels across these approaches.

Furthermore, our study used multi-center data with a relatively higher prevalence of poor-quality examinations—an area where Ponsiglione et al. suggested further research was needed, as their sample mostly included good-quality exams. Despite methodological and sample differences, Ponsiglione et al. found that PI-

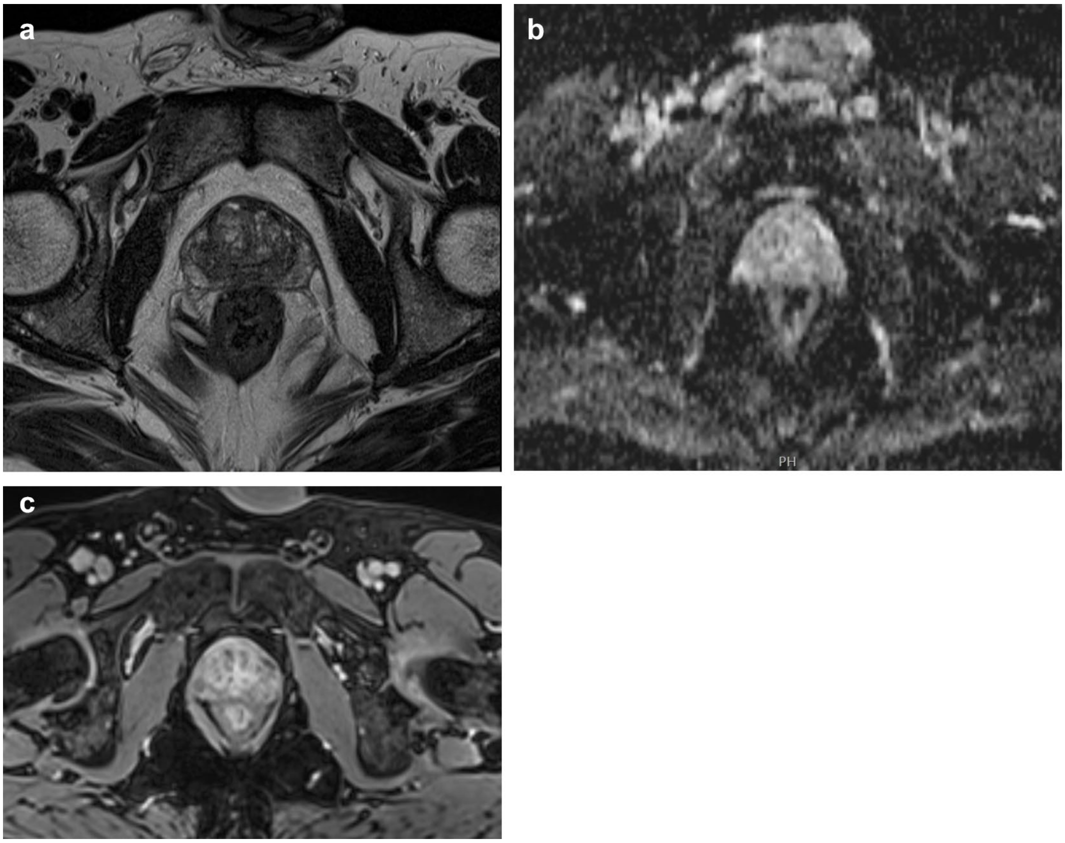

Fig. 5. A 62-year-old man who underwent multi-parametric prostate MRI at 1.5 T. (**a**) On the axial T2-weighted image, there is prominent noise with a relatively low signal-to-noise ratio. Anatomical structures are mostly delineated, and no artifacts are present; however, the signal-to-noise ratio remains low on both coronal and sagittal T2 W images (not shown). (**b**) The anatomically matched axial apparent diffusion coefficient (ADC) map is free of artifacts, but the contrast and signal-to-noise ratio are relatively low, though delineation of the peripheral and transitional zones is possible. (**c**) Vessels are clearly delineated, with no visible artifacts on the dynamic-contrast-enhanced image. Reader 1, 4, and 5 assigned a Prostate Imaging Quality (PI-QUAL) V2 score of 2, while reader 2 and 3 assigned a PI-QUAL V2 score of 1 for this case. For PI-QUAL V1, reader 1 and 4 assigned a score of 4, reader 2 and 5 assigned a score of 3, and reader 3 assigned a score of 2.

QUAL V2 was moderately reliable for image quality assessment. In contrast, our study found almost perfect agreement levels among basic readers using PI-QUAL V2, indicating stronger performance. Additionally, we directly compared PI-QUAL V1 and V2, demonstrating improvements in inter-rater agreement with the updated version.

A recent study by Lee et al. compared PI-QUAL V1 and V2 on 350 multiparametric MRI scans, using scores assigned by two expert uroradiologists[18]. In contrast to our findings, the authors reported substantial agreement for overall PI-QUAL V1 ($\kappa = 0.64$) and moderate agreement for PI-QUAL V2 ($\kappa = 0.54$). As expected, most examinations in their study were assigned higher quality scores, resulting in a skewed distribution. The authors used kappa scores instead of Gwet's AC1 for their analysis. Kappa scores are known to decrease when one category dominates, as they rely on expected agreement under marginal distributions. In contrast, Gwet's AC1 is more robust to such issues, as it adjusts for chance agreement in a manner less sensitive to prevalence effects and marginal imbalances. This robustness likely explains the relatively lower agreement levels reported in their study, despite the readings being performed by expert radiologists.

Our study has several limitations. First, we did not evaluate how image quality impacts readers' diagnostic performance in detecting csPCa. Assessing this relationship retrospectively across a heterogeneous patient population poses challenges, as factors such as variations in patient demographics, tumor characteristics, and potential inconsistencies in image acquisition and interpretation complicate the analysis, making definitive conclusions difficult. However, the primary objective of this study was to assess inter-rater agreement for PI-QUAL V2 and compare it with V1 among basic readers using a multi-center dataset with heterogeneous image quality, rather than focusing on the impact of image quality on diagnostic performance.

Second, although our dataset is multi center and multi scanner, all examinations were acquired on Siemens platforms. While inclusion of both 1.5 T and 3 T magnets, diverse coil configurations (8, 16, and 18 channel surface arrays), and center specific sequence parameters introduced some technical variability, vendor-specific reconstruction pipelines may still influence signal to noise characteristics and artifact profiles. To enhance generalizability, future prospective studies should purposefully incorporate scanners from additional manufacturers—particularly GE, Philips, Canon, and United Imaging—and evaluate PI QUAL V2 reproducibility across mixed vendor cohorts.

Third, although our results demonstrate that PI QUAL V2 yields higher and more uniform inter reader agreement than V1, we did not investigate whether that improvement translates into better diagnostic performance or patient outcomes. Establishing such a link would require (i) standardized biopsy results or longitudinal follow up as ground truth, (ii) a substantially larger, prospectively collected cohort to ensure adequate power, and (iii) multivariable modeling to account for reader experience and tumor prevalence. A systematic review by Stabile et al. showed that lower MRI quality correlates with reduced csPCa detection on prostate MRI, underscoring the clinical relevance of image quality. Accordingly, we view the current study as a critical methodological step—providing robust evidence of reproducibility—that paves the way for prospective, multi-vendor trials designed to test whether improved inter-reader agreement with PI-QUAL V2 translates into higher cancer detection rates and better clinical outcomes.

In conclusion, our study found that five basic prostate readers achieved almost perfect agreement in evaluating prostate image quality using the PI-QUAL V2 score on multi-center data. Furthermore, inter-rater agreement among readers was higher for PI-QUAL V2 compared to PI-QUAL V1, indicating that the revised version offers greater robustness and reliability for clinical use.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 20 December 2024; Accepted: 16 May 2025
Published online: 02 July 2025

References

1. Ahmed, H. U. et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389**, 815–822 (2017).
2. Bjurlin, M. A. et al. Update of the standard operating procedure on the use of multiparametric magnetic resonance imaging for the diagnosis, staging and management of prostate Cancer. *J. Urol.* **203**, 706–712 (2020).
3. EAU Guidelines on Prostate Cancer - Uroweb. *Uroweb - Eur. Association Urology* https://uroweb.org/guidelines/prostate-cancer
4. de Rooij, M. et al. Focus on the quality of prostate multiparametric magnetic resonance imaging: synopsis of the ESUR/ESUI recommendations on quality assessment and interpretation of images and radiologists' training. *Eur. Urol.* **78**, 483–485 (2020).
5. de Rooij, M. et al. ESUR/ESUI consensus statements on multi-parametric MRI for the detection of clinically significant prostate cancer: quality requirements for image acquisition, interpretation and radiologists' training. *Eur. Radiol.* **30**, 5404–5416 (2020).
6. Esses, S. J., Taneja, S. S. & Rosenkrantz, A. B. Imaging facilities' adherence to PI-RADS v2 minimum technical standards for the performance of prostate MRI. *Acad. Radiol.* **25**, 188–195 (2018).
7. Burn, P. R. et al. A multicentre assessment of prostate MRI quality and compliance with UK and international standards. *Clinical Radiology* **74**, (2019). 894.e19-894.e25.
8. Sackett, J. et al. Quality of prostate MRI: is the PI-RADS standard sufficient?? *Acad. Radiol.* **28**, 199–207 (2021).
9. A trial looking at using MRI to help diagnose prostate cancer (PRECISION). (2016). https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-trial-looking-at-using-mri-to-help-diagnose-prostate-cancer-precision
10. Giganti, F. et al. Prostate imaging quality (PI-QUAL): A new quality control scoring system for multiparametric magnetic resonance imaging of the prostate from the PRECISION trial. *Eur. Urol. Oncol.* **3**, 615–619 (2020).
11. Giganti, F. et al. Inter-reader agreement of the PI-QUAL score for prostate MRI quality in the neurosafe PROOF trial. *Eur. Radiol.* **32**, 879–889 (2022).
12. Karanasios, E., Caglic, I., Zawaideh, J. P. & Barrett, T. Prostate MRI quality: clinical impact of the PI-QUAL score in prostate cancer diagnostic work-up. *BJR* **95**, 20211372 (2022).
13. Stabile, A. et al. Factors influencing variability in the performance of multiparametric magnetic resonance imaging in detecting clinically significant prostate cancer: A systematic literature review. *Eur. Urol. Oncol.* **3**, 145–167 (2020).
14. Giganti, F. et al. Understanding PI-QUAL for prostate MRI quality: a practical primer for radiologists. *Insights into Imaging.* **12**, 59 (2021).
15. de Rooij, M. et al. PI-QUAL version 2: an update of a standardised scoring system for the assessment of image quality of prostate MRI. *Eur. Radiol.* **34**, 7068–7079 (2024).
16. Basar, Y. et al. Inter-reader agreement of the prostate imaging quality (PI-QUAL) score for basic readers in prostate MRI: A multi-center study. *Eur. J. Radiol.* **165**, 110923 (2023).
17. Ponsiglione, A. et al. PI-QUAL version 2: A Multi-Reader reproducibility study on multiparametric MRI from a tertiary referral center. *Eur. J. Radiol.* **181**, 111716 (2024).
18. Lee, K. L. et al. PI-QUAL version 2 image quality categorisation and inter-reader agreement compared to version 1. *Eur. Radiol.* https://doi.org/10.1007/s00330-024-11233-1 (2024).

Author contributions

Each author has made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work. Each author has approved the submitted version (and any substantially modified version that involves the author's contribution to the study). Each author has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. Please find the detailed list of each author's contribution to the present work below. M.O.: Conceptualization, Data curation; S.S.: Data curation; M.E.S.: Formal analysis; D.A.: Writing, Conceptualization; M.S.K.: Formal analysis; Y.B.: Data curation; N.D.: Data curation; S.K.: Data curation; A.A.: Data curation; I.O.: Conceptualization, Critical review; E.K.: Conceptualization, Critical review.

Declarations

Competing interests
D.A. is the CEO and co-founder of Hevi AI Health Tech and M.S.K. is employed as a medical AI engineer at the same company.None of Hevi AI's products were used or mentioned in the current work. Furthermore, this paper did not use any commercially available deep-learning software.The other authors have declared no conflicts of interest. None of Hevi AI's products were used or mentioned in the current work. Furthermore, this paper did not use any commercially available deep-learning software. The other authors have declared no conflicts of interest.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-02973-1.

Correspondence and requests for materials should be addressed to D.A.