

Article

Repeatability of Semi-Quantitative and Volumetric Features from Artificial-Intelligence-Guided Lesion Segmentation on ^{18}F -DCFPyL PSMA-PET/CT Images: Results from a Test-Retest Cohort

Md Zobaer Islam ¹, Timothy G. Perk ² , Amy Weisman ² , Mark C. Markowski ³ , Kenneth J. Pienta ³ , Young E. Whang ⁴ , Matthew I. Milowsky ⁴ , Martin G. Pomper ¹ , Nicholas Wisniewski ⁵, Ralph A. Bundschuh ^{6,7}, Rudolf A. Werner ⁸ , Michael A. Gorin ⁹  and Steven P. Rowe ^{1,*}

¹ Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

² AIQ Solutions, Madison, WI 53717, USA

³ Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

⁴ Division of Oncology, Department of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA

⁵ Department of Radiology, University of North Carolina, Chapel Hill, NC 27599, USA

⁶ Department of Nuclear Medicine, Universitätsklinikum Carl Gustav Carus at the TU Dresden, 01307 Dresden, Germany

⁷ German Cancer Consortium (DKTK), Partner Site Dresden, 01307 Dresden, Germany

⁸ Department of Nuclear Medicine, LMU University Hospital, LMU Munich, 80539 Munich, Germany

⁹ Milton and Carroll Petrie Department of Urology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

* Correspondence: steven.rowe@utsouthwestern.edu; Tel.: +1-734-717-0909

Simple Summary

To date, the test–retest repeatability of lesion-level features derived from artificial intelligence (AI)-guided prostate-specific membrane antigen (PSMA)-PET lesion segmentation has not been systematically assessed. One reason for that is the lack of available data. We demonstrate that a unique test–retest dataset of PSMA-PET scans, i.e., paired scans of patients with metastatic prostate cancer obtained within one week of each other, provides a test bed for AI algorithms to demonstrate how repeatably they can identify and delineate tumors. The methodology we describe is a new means of assessing the validity of AI algorithms.

Abstract

Objectives: This study evaluated the test–retest repeatability of semi-quantitative and volumetric features derived from artificial intelligence (AI)-assisted lesion segmentation on ^{18}F -DCFPyL Prostate Specific Membrane Antigen (PSMA)-PET/CT imaging of patients with prostate cancer (PCa). Specifically, we assessed the reliability of maximum, minimum and total standardized uptake values (SUV_{max} , SUV_{mean} , $\text{SUV}_{\text{total}}$) and lesion volume measurements across varying lesion sizes and explored the implications of variability for clinical decision-making. **Methods:** We analyzed ^{18}F -DCFPyL PSMA-PET/CT images from 22 patients with metastatic PCa. Lesion segmentation was performed using the AI-guided TRAQinform IQ technology, followed by a manual review to eliminate potential false-positive sites of uptake. Lesion-level test–retest repeatability was evaluated using 95% limits of agreement (LOA), intra-class correlation coefficient (ICC), within-subject coefficient of variation (wCOV) and Bland–Altman analysis for SUV and volumetric parameters. Lesions were stratified by size ($>1\text{ cm}^3$ and $>1.5\text{ cm}^3$) to assess the impact of lesion volume cut-offs on measurement variability. **Results:** A total of 297 lesions were analyzed,



Academic Editors: Emilio Quaia and James M. Mountz

Received: 14 December 2025

Revised: 12 February 2026

Accepted: 6 March 2026

Published: 11 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

including 191 lesions $> 1 \text{ cm}^3$ and 161 lesions $> 1.5 \text{ cm}^3$. Test–retest variability was higher in smaller lesions, with narrower LOA and lower wCOV for larger lesions. SUV_{max} and SUV_{mean} exhibited lower variability than $\text{SUV}_{\text{total}}$ and lesion volume. The 95% LOA for SUV_{max} ranged from -33.81% to $+38.02\%$ for all lesions, improving to -31.82% to $+31.01\%$ for lesions $> 1.5 \text{ cm}^3$. Similar trends were observed for SUV_{mean} , $\text{SUV}_{\text{total}}$, and volume. Bland–Altman plots confirmed reduced variability in larger lesions, with no significant systematic bias. **Conclusions:** The test–retest repeatability of AI-assisted PSMA-PET/CT features varies by feature type, with semi-quantitative features demonstrating improved repeatability relative to volumetric features. Additionally, repeatability is influenced by lesion size, with larger lesions exhibiting greater reliability. These findings highlight the importance of lesion size-dependent thresholds in response assessment and variability-aware feature selection in prognostic models. Current algorithms may be better optimized for larger lesions and higher volumes of disease, with limitations remaining in the robust detection and segmentation of smaller/more subtle lesions.

Keywords: PSMA-PET/CT; test–retest; PET features; artificial intelligence

1. Introduction

Prostate cancer (PCa) is one of the most commonly diagnosed malignancies among men worldwide and remains a major cause of cancer-related morbidity and mortality, particularly in advanced and metastatic stages [1,2]. Accurate detection, staging, and longitudinal assessment of disease burden are essential for guiding treatment selection, evaluating therapeutic response, and improving patient outcomes [3]. Prostate-specific membrane antigen (PSMA)-targeted positron emission tomography/computed tomography (PET/CT) imaging has emerged as a highly sensitive and specific modality for detecting and characterizing PCa, particularly in metastatic and recurrent disease settings [4–6]. By leveraging the high expression of PSMA in PCa cells, PSMA-PET/CT enables precise localization of tumor lesions, aiding in disease staging, treatment planning, radiation dose estimation, and therapeutic response assessment [7–9]. Despite these advantages, the reliability and reproducibility of quantitative and volumetric metrics derived from PSMA-PET/CT images remain critical but uncertain factors in ensuring accurate disease monitoring and therapy evaluation [10]. Understanding the test–retest repeatability of those features is essential to establishing their robustness for clinical and research applications.

One of the major challenges in PSMA-PET/CT analysis is the limited availability of histologic ground truth for metastatic tumor volume, making it difficult to biologically validate segmentation results. Traditionally, semi-automated or manual delineation of tumor lesions by expert radiologists has been employed for segmentation tasks; however, inter-reader and intra-reader variability pose concerns regarding the consistency of those methods [10]. In such scenarios, test–retest repeatability offers a pseudo-standard for evaluating the reliability of features extracted from regions of interest (ROI) by assessing how consistently features are reproduced in sequential scans under similar conditions. Indeed, when there can be no histologic ground-state truth, repeatability may be the most important aspect of a segmentation algorithm to validate it for clinical use.

Prior studies have investigated the test–retest repeatability of quantitative features from PSMA-targeted ^{68}Ga -PSMA and ^{18}F -DCFPyL PET/CT datasets, utilizing semi-automated and/or manual segmentation [11,12]. Those studies have demonstrated, using statistical metrics, that while commonly used semi-quantitative parameters such as maximum standardized uptake value (SUV_{max}) and mean standardized uptake value (SUV_{mean})

exhibit relatively high repeatability, volumetric features such as PSMA-tumor volume and total lesion PSMA tend to show greater variability. Additionally, radiotracer uptake intensity has been correlated with repeatability, with higher uptake lesions demonstrating better, more robust repeatability [12]. However, when high-frequency radiomic features were analyzed, a substantial proportion exhibited poor repeatability, suggesting that feature robustness varies significantly depending on the type of features considered [13]. Given the increasing adoption of radiomics and artificial intelligence (AI)-derived features in oncologic imaging, establishing the reproducibility of such extracted features is crucial but unrealized.

With the advent of AI-driven methodologies in medical imaging, deep learning-based segmentation techniques have gained traction in automating lesion detection and delineation [14–16]. AI-guided segmentation offers several advantages, including reduction in observer bias, improved efficiency, and potential for enhanced reproducibility. Nonetheless, lesion-level test–retest repeatability of AI-assisted segmentation features has not been systematically analyzed in the context of PSMA-PET/CT imaging or PET imaging, more broadly. Understanding whether AI-assisted segmentation approaches can produce robust and repeatable quantitative and volumetric features is critical for their translation into routine clinical practice. In this context, the key contribution of the present study is a technical validation of AI-guided segmentation by assessing lesion-level test–retest repeatability of commonly used semi-quantitative and volumetric features on ^{18}F -DCFPyL PSMA-PET/CT images. By focusing on repeatability under conditions where histologic ground truth is unavailable, this work provides a practical framework for evaluating the robustness of AI-guided segmentation outputs. The findings of this study have significant implications for clinical practice, as repeatable AI-extracted features could facilitate the development of reliable imaging biomarkers for disease monitoring, risk stratification, and treatment response assessment.

2. Materials and Methods

2.1. Study Design

This study utilized ^{18}F -DCFPyL PSMA-PET/CT images in DICOM format from 22 patients diagnosed with metastatic PCa. All patients were originally accrued on an institutional review board-approved prospective protocol. PET/CT imaging was performed using a Siemens Biograph 128-slice mCT scanner (Siemens Healthineers, Erlangen, Germany) under standardized acquisition protocols to ensure reproducibility. Each patient underwent two PET/CT scans within a range of 1–7 days to assess test–retest repeatability, with no tumor-specific therapy administered between scans. Imaging was conducted approximately 60 min post-injection (range: 57–63 min) following intravenous administration of 322.4 MBq (first scan) and 323.6 MBq (second scan) of ^{18}F -DCFPyL. PET data were acquired from mid-thigh to skull vertex over 6–8 bed positions (3 min per position), alongside a low-dose CT scan for attenuation correction and anatomical localization. Image reconstruction utilized the manufacturer’s ordered-subset expectation maximization algorithm, incorporating scatter and attenuation corrections. The study was registered at ClinicalTrials.gov (NCT03793543) and conducted under a United States Food and Drug Administration (FDA) Investigational New Drug Application (IND121064) as ^{18}F -DCFPyL was not FDA-approved at the time. Detailed patient characteristics including age, prior treatments, and disease burden, as well as imaging acquisition protocols, have been previously published [12,17]. A graphical overview of the workflow for this study is shown in Figure 1.

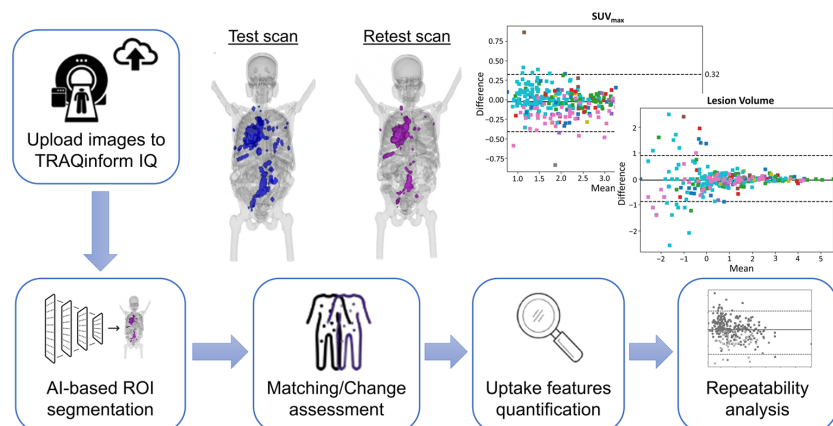


Figure 1. Workflow of the study from image upload to feature extraction and repeatability analysis.

2.2. AI-Guided Segmentation

Lesion segmentation from the images was performed using AI-guided TRAQinform IQ version 1.9 technology (AIQ Solutions, Madison, WI, USA). This software employed an automated lesion detection and segmentation algorithm based on a Retina U-Net architecture, applied to both baseline and follow-up scans [18]. Retina U-Net is a deep learning-based model that combines the RetinaNet object detection framework [19] with a U-Net segmentation architecture [20] to enhance lesion detection accuracy. Unlike conventional semantic segmentation approaches, Retina U-Net leverages a feature pyramid network [21] that enables multi-scale object detection while simultaneously integrating full-resolution segmentation supervision. That architecture refines object-level lesion detection by incorporating both pixel-wise and object-wise contextual information, improving accuracy in challenging medical imaging tasks [22]. The Retina U-Net was trained on a heterogeneous, multi-institutional patient dataset and evaluated using an external hold-out validation cohort that included both ^{68}Ga -PSMA and ^{18}F -labeled PSMA images. The software operated as a cloud-based service, where PET/CT images in DICOM format were uploaded to the platform and segmentation results were returned. The automated results were further refined through manual review by a quality assurance team led by a nuclear medicine technologist with over 22 years of experience to eliminate false-positive ROIs when these were located in healthy organs. A representative example of the AI-guided segmentation results, after expert refinement, is shown on test–retest maximum intensity projection images in Figure 2.

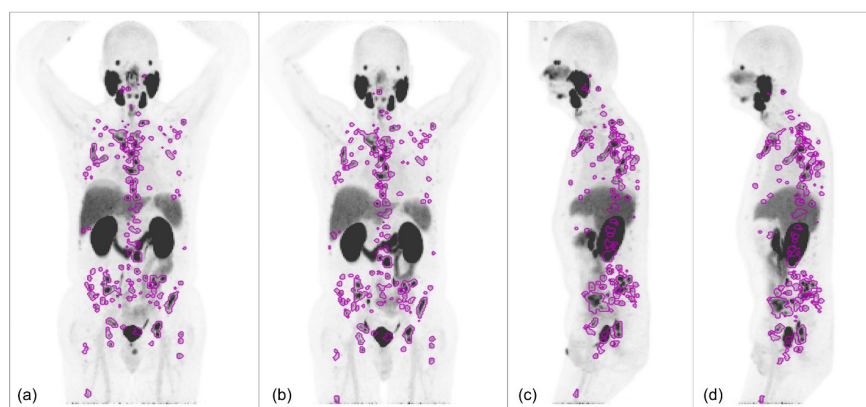


Figure 2. Test–retest ^{18}F -DCFPyL PSMA-PET/CT maximum intensity projection images from a representative patient with metastatic prostate cancer. The images show AI-guided lesion segmentation (magenta outlines) on (a) test coronal, (b) retest coronal, (c) test sagittal, and (d) retest sagittal views. The retest scan was performed within 7 days of the initial scan.

2.3. Feature Extraction

To track lesions over time, a registration-based method was implemented to match ROIs across the two time-points. That process involved first aligning the images from both scans and then determining which lesions corresponded between baseline and follow-up scans. This registration-based lesion matching approach has been previously validated and benchmarked against inter-reader variability, with no significant differences observed for precision, recall, F1-score, and the number of differences [23]. From the segmented and matched ROIs, the software extracted two semi-quantitative features (SUV_{max} and SUV_{mean}) and two volumetric features (volume and SUV_{total}) for each individual lesion at both time points. SUV_{max} represented the highest standardized uptake value (SUV) within the ROI, while SUV_{mean} indicated the average SUV within the ROI. Volume refers to the total size of the ROI measured in cm^3 , and SUV_{total} is determined by summing the SUVs of all voxels within the ROI and multiplying by the voxel volume.

2.4. Repeatability Analysis

To assess the test–retest repeatability of extracted features, multiple repeatability metrics including limits of agreement (LOA), intra-class correlation coefficients (ICC) [24] and within-subject coefficients of variations (wCOV) were calculated for all lesions. Additionally, Bland–Altman analysis was conducted to evaluate the limits of agreement of measurements and provide a visual representation of measurement differences and potential trends in variability [25]. Since SUV-based metrics often exhibit a skewed distribution, a natural log transformation was applied to the feature quantities to enhance statistical robustness before calculating the 95% limits of agreement (LOA) for the ratio between test and retest measurements. The final LOA values were then converted back using the formula:

$$95\% \text{ LOA} = \left(e^{(\bar{d}-1.96\sigma)}, e^{(\bar{d}+1.96\sigma)} \right),$$

where bias $\bar{d}$ represented the mean ratio between test and retest measurements and σ was the standard deviation of the mean ratios [26]. Since many patients had multiple lesions with diverse radiotracer uptake levels, the calculation of $\bar{d}$ and σ followed the methodology designed for estimating limits of agreement in scenarios involving multiple observations per individual, where the true value may vary [27].

To assess the impact of lesion size on repeatability, additional analyses were performed by excluding ROIs below predefined volumetric thresholds. Specifically, repeatability metrics were recalculated, and Bland–Altman plots were recreated after excluding lesions smaller than 1 cm^3 and 1.5 cm^3 , as smaller lesions are more susceptible to segmentation variability. These thresholds were selected based on prior PET imaging literature, where similar lesion size cut-offs have been used when evaluating feature repeatability and disease progression [26,28]. All statistical analyses to evaluate repeatability in this study were performed with Python version 3.9.

3. Results

3.1. Repeatability Across All Lesions

A total of 297 matching lesions of varying volumes were delineated using TRAQinform IQ software, with 191 lesions having a volume greater than 1 cm^3 and 161 exceeding 1.5 cm^3 . To evaluate the test–retest repeatability of SUV_{max} , SUV_{mean} , SUV_{total} , and volume, 95% limits of agreement, ICC and wCOV were computed. Table 1 presents these metrics across all lesions, as well as for lesions exceeding 1 cm^3 and 1.5 cm^3 , respectively. The 95% LOA for each uptake parameter, as presented in Table 1 in percentage, indicates the expected range within which the percentage differences between test and retest measurements lie

for 95% of the lesions. Those values account for inherent sources of variability in PET-based lesion quantification, including imaging noise, segmentation inconsistencies, and physiological fluctuations.

Table 1. Repeatability metrics of lesion uptake features for (A) all lesions, (B) lesions larger than 1 cm³ and (C) lesions larger than 1.5 cm³.

(A) All lesions (number of lesions: 297)				
	SUV _{max}	SUV _{mean}	SUV _{total}	Lesion Volume
Lower LOA (%)	−33.81	−25.78	−61.34	−58.62
Upper LOA (%)	38.02	24.10	142.36	145.89
ICC	0.973	0.960	0.972	0.996
wCOV	9.13	9.42	22.22	63.67
(B) Lesions having volume > 1 cm ³ (number of lesions: 191)				
Lower LOA (%)	−31.72	−24.27	−35.15	−33.07
Upper LOA (%)	32.31	23.38	38.48	43.83
ICC	0.972	0.958	0.974	0.996
wCOV	6.88	7.82	6.08	12.27
(C) Lesions having volume > 1.5 cm ³ (number of lesions: 161)				
Lower LOA (%)	−31.82	−25.74	−34.88	−31.13
Upper LOA (%)	31.01	24.26	40.54	44.31
ICC	0.971	0.949	0.972	0.995
wCOV	6.50	7.90	5.62	10.34

3.2. Effects of Lesion Size

The results demonstrate that lesion size plays a crucial role in test–retest agreement, with smaller lesions exhibiting greater variability. As shown in Table 1, excluding lesions smaller than 1 cm³ results in narrower LOA ranges and lower wCOV values, indicating improved measurement repeatability. This effect becomes more pronounced when considering only lesions greater than 1.5 cm³, where the LOA range further contracts and wCOV is further reduced in most of the features, indicating increased reliability of SUV and volume measurements in larger lesions.

3.3. Bland–Altman Analysis

Bland–Altman plots of the selected features are presented in Figure 3 on log-log scales for all lesion ROIs, in Figure 4 for ROIs with volume > 1 cm³, and in Figure 5 for ROIs with volume > 1.5 cm³, with color coding corresponding to individual patients. In those plots, the mean difference (bias) is close to zero across all uptake parameters, suggesting no significant systematic overestimation or underestimation between test and retest scans. The spread of the LOAs indicates variability in repeat measurements. Smaller lesions show a wider spread of differences (higher variability) than larger lesions, suggesting increased measurement uncertainty in smaller lesions. For SUV_{max} and SUV_{mean}, the differences between test and retest scans remain relatively more stable across lesion sizes, though some heteroscedasticity is observed, with greater variability in smaller lesions. In contrast, SUV_{total} and volume exhibit more pronounced spread in smaller lesions, which suggests that volume-dependent effects influence test–retest variability for these lesions that have

been segmented via AI assistance. In Figures 4 and 5, the LOAs are narrower, further confirming that repeatability improves with larger lesion sizes.

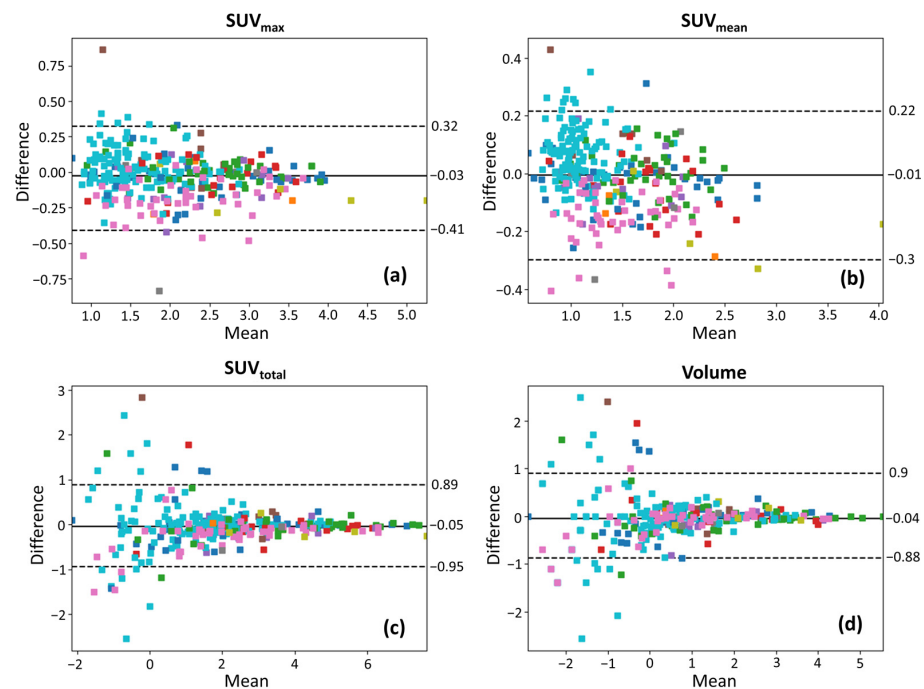


Figure 3. Bland–Altman plots on log-log scales, illustrating the test–retest variability of (a) SUV_{max} , (b) SUV_{mean} , (c) SUV_{total} , and (d) lesion volume for all lesions. The solid line represents the mean difference, and dashed lines indicate the 95% limits of agreement. Data points are color-coded to represent different patient IDs.

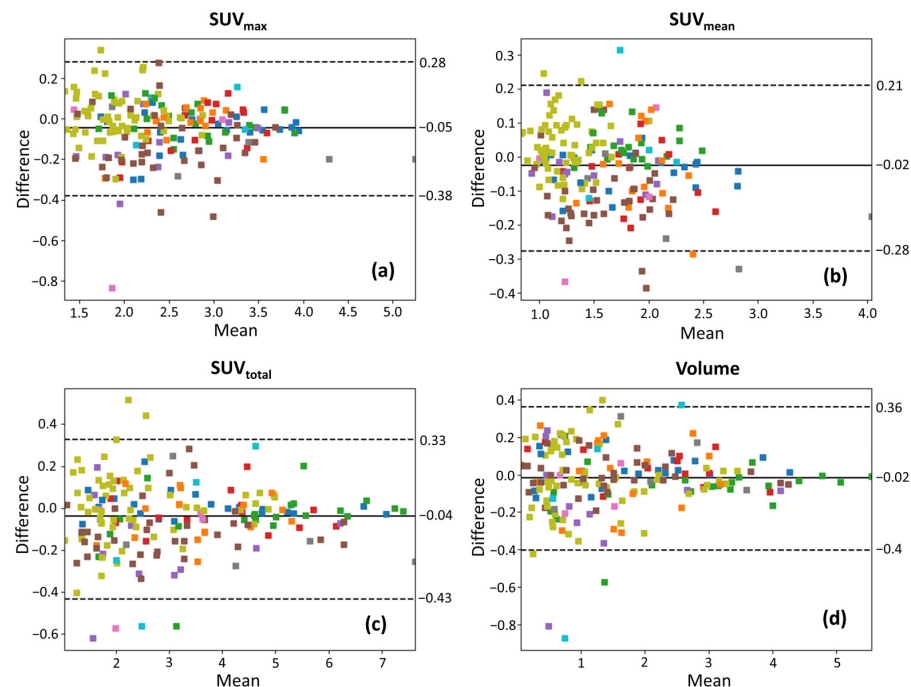


Figure 4. Bland–Altman plots on log-log scales, illustrating the test–retest variability of (a) SUV_{max} , (b) SUV_{mean} , (c) SUV_{total} , and (d) lesion volume for lesions larger than 1 cm³. The solid line represents the mean difference, and dashed lines indicate the 95% limits of agreement. Data points are color-coded to represent different patient IDs.

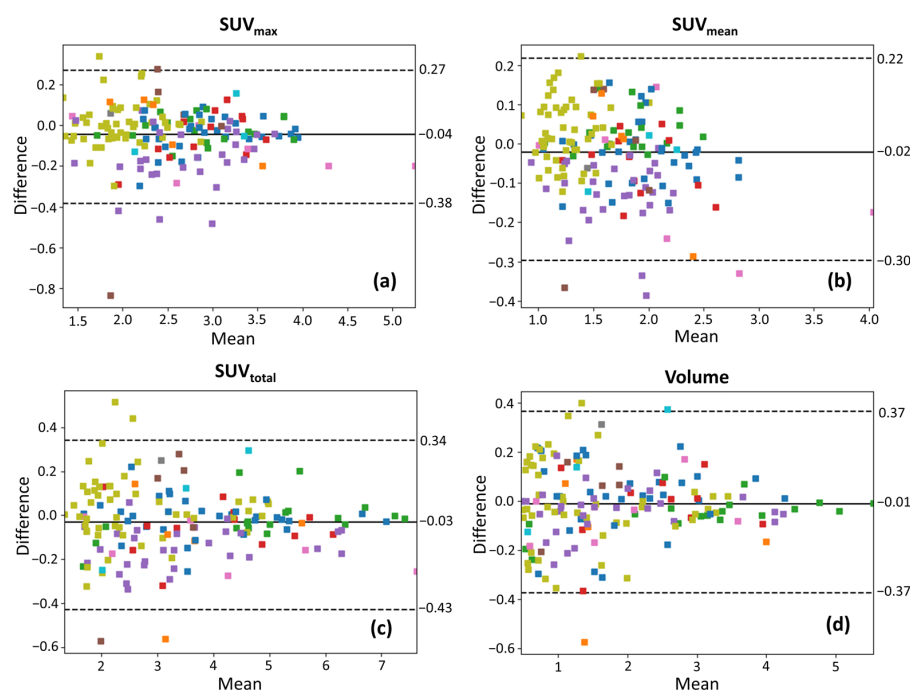


Figure 5. Bland–Altman plots on log-log scales, illustrating the test–retest variability of (a) SUV_{max} , (b) SUV_{mean} , (c) SUV_{total} , and (d) lesion volume for lesions larger than 1.5 cm^3 . The solid line represents the mean difference, and dashed lines indicate the 95% limits of agreement. Data points are color-coded to represent different patient IDs.

4. Discussion

4.1. Semi-Quantitative vs. Volumetric Feature

This study demonstrates that SUV_{max} , SUV_{mean} , SUV_{total} , and lesion volume exhibit test–retest variations of approximately $\pm 30\text{--}40\%$ (Table 1). This suggests that small changes in these metrics may fall within normal variability rather than reflecting true treatment response. ICC remains consistently high (>0.94) across all lesion groups and features in Table 1, suggesting that despite variability in individual lesion measurements, there is a strong overall agreement between test and retest scans. Compared to SUV_{max} and SUV_{mean} , SUV_{total} and lesion volume exhibit higher wCOV and broader LOA ranges, which indicates that volumetric features are less repeatable than semi-quantitative SUV features.

Bland–Altman plots reveal distinct patient-dependent biases in SUV_{max} and SUV_{mean} . For example, in Figure 3a,b, cyan-colored points skew positive while magenta-colored points skew negative with respect to the mean difference lines, suggesting systemic differences between patients. Potential explanations for these biases include variations in administered dose, tracer uptake kinetics, patient positioning, and physiological factors. However, patient-dependent biases appear to primarily influence SUV_{max} and SUV_{mean} but are less pronounced in SUV_{total} and volume, suggesting that volumetric parameters may be less influenced by these sources of variability. This implies that SUV_{total} variability may arise from two distinct sources: systemic PET measurement fluctuations (affecting SUV_{mean}) and segmentation-derived volume differences. Future quantitative analysis will be needed to quantify the contributions of these two factors to SUV_{total} variability.

4.2. Lesion Size Dependence of Repeatability

Small lesions exhibit a unique pattern in volumetric measurement variability in Bland–Altman plots, evident in Figure 3c,d, where the variability increases abruptly, rather than gradually, as the lesion size decreases. This indicates a potential threshold effect, likely due to voxel quantization. Because small lesions contain a limited number of voxels, volume

estimation becomes highly sensitive to even slight variations in segmentation. This effect introduces a “noise floor”, where minor changes in segmentation lead to disproportionately large differences in measured volume. Future work could involve reporting lesion sizes in both voxel count and physical units, as well as categorizing lesions into discrete voxel-based size bins, to provide a more precise evaluation of segmentation repeatability across different lesion sizes.

This study further shows that smaller lesions exhibit greater test–retest variability, making SUV-derived features from these lesions less stable. For widely metastatic patients undergoing systemic therapy and for whom response assessment is desired, machine learning models trained on PET imaging should consider excluding small lesions or prioritizing larger, more stable lesions, such as those exceeding 1.5 cm³, to enhance robustness. Additionally, feature selection techniques that account for variability should be incorporated to ensure that only highly repeatable features contribute to model predictions. For patients with low-volume metastatic disease, whose tumors may be small and/or subtle, existing approaches may need to be significantly refined to ensure repeatability and robustness, given the poor outcomes that can arise from failing to effectively treat all of the lesions in such patients [29].

4.3. Comparison with Manual Segmentation

Our AI-guided segmentation approach resulted in repeatability that was comparable to our previously reported manual segmentation-based analysis on the same test–retest cohort [12,17]. For example, the wCOV for SUV_{max} and SUV_{mean} in the current study, considering all the lesions, are 9.13% and 9.42%, similar to previously reported manual results (12.1% and 7.3%) [12]. While repeatability is comparable, the AI method offers clear advantages in processing efficiency and eliminates inter- and intra-reader variability.

4.4. Implications for Response Assessment

Quantitative PET imaging provides essential biomarkers for tumor detection, treatment response assessment, and prognostication [30]. However, inherent test–retest variability in PET-derived parameters must be carefully considered when defining thresholds for response classification. The observed negative percentage LOAs (Table 1) suggest that minor reductions in these values may reflect normal variability rather than actual disease regression. SUV_{total} and lesion volume show the highest test–retest variability in small lesions, which can affect prognostic models that rely on tumor burden quantification. To improve the reliability of volumetric PET biomarkers, excluding lesions below a threshold (e.g., 1.5 cm³) could enhance consistency. Standardizing PET-based biomarkers and refining response assessment frameworks like PERCIST or RECIP to integrate lesion size-specific adjustments will be critical for improving the reliability of PET imaging in both clinical and research settings [31,32]. In the context of metastatic disease, where a histologic ground truth is unlikely to ever be realistically available, lesion segmentation algorithms might best be judged by their repeatability. The more repeatable the algorithm’s semi-quantitative and volumetric outputs, the more robust the predictive and prognostic biomarkers derived from that algorithm should be.

Furthermore, the variability in the robustness of AI-assisted segmentation based on lesion size suggests that the current algorithm is best suited for delineating prominent disease sites in patients who are appropriate candidates for systemic therapy. However, patients with lower disease burden and more subtle metastatic lesions may require an algorithm optimized for detecting smaller lesions with lower SUV parameters. Interestingly, our findings align with those from a recent meta-analysis of ¹⁸F-FDG-PET studies [33], which demonstrated that lesions with higher uptake values showed better repeatability in

terms of percent fluctuation across SUV metrics. This is consistent with our observations in PSMA-PET, where relative variability (as measured by wCOV) decreased with increasing SUV and lesion volume. These findings reinforce the broader principle that small or low-uptake lesions are more susceptible to biological and technical variability, regardless of the radiotracer used.

4.5. Limitations

The limitations of this study include that it is a post hoc analysis of a prospectively acquired dataset as opposed to a prospective and powered study to meet a defined endpoint. Nonetheless, this analysis was carried out with the largest PSMA-PET/CT test–retest dataset yet reported and obtaining larger datasets may be cost-prohibitive or impractical. Further, there are parameters yet to be explored such as whether there are metastasis–location-dependent variations in the repeatability of semi-quantitative or volumetric parameters. In addition, repeatability metrics were computed at the lesion level rather than the patient level, which may underrepresent patient-level biological and technical variability. Furthermore, while repeatability is crucial, it does not necessarily indicate that a feature is always sensitive to treatment-induced changes. Prior research has introduced the concept of “response-to-repeatability” [34], highlighting the need to assess whether highly repeatable PET features are also effective for monitoring treatment response. Future studies in PSMA-PET could explore this relationship to refine PET-based biomarkers for clinical decision-making.

5. Conclusions

This study emphasizes the critical need for standardization in PSMA-PET imaging protocols to minimize variability and improve measurement reproducibility, even when using AI-driven algorithms. The results highlight the significance of accounting for test–retest variability, particularly when using features derived from AI-driven lesion segmentation. While volumetric PSMA-PET metrics exhibit substantial fluctuations between repeat scans, even in the absence of biological changes, larger lesions tend to offer more stable and reproducible measurements. This reinforces the need for size-dependent response assessment strategies. Future research should focus on improving AI-driven segmentation and feature extraction models by integrating uncertainty quantification methods that can identify and mitigate the impact of unstable features. Such models could generate voxel-wise or lesion-level confidence maps, allowing clinicians to selectively trust high-confidence segmentations while flagging ambiguous or variable regions for manual review or exclusion from downstream analysis. The use of multimodal imaging, such as PET/MRI fusion, may further enhance the robustness of PET-based biomarkers by providing complementary anatomical and functional insights. Additionally, developing adaptive thresholding techniques that adjust for lesion size and inherent measurement variability could refine response assessment criteria, making PET imaging more reliable for precision oncology applications.

Author Contributions: Conceptualization, M.Z.I. and S.P.R.; methodology, M.Z.I., T.G.P., A.W. and S.P.R.; software, T.G.P. and A.W.; formal analysis, T.G.P., A.W., N.W. and R.A.B.; investigation, M.C.M., K.J.P., Y.E.W., M.I.M., R.A.W. and M.A.G.; data curation, M.Z.I., T.G.P., A.W., R.A.B. and R.A.W.; writing—original draft preparation, M.Z.I. and S.P.R.; writing—review and editing, T.G.P., A.W., M.C.M., K.J.P., Y.E.W., M.I.M., M.G.P., N.W., R.A.B., R.A.W. and M.A.G.; supervision, M.G.P., R.A.W., M.A.G. and S.P.R.; funding acquisition, M.G.P. and S.P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a Prostate Cancer Foundation Young Investigator Award.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Johns Hopkins (IRB00174393 on 6 August 2020). The study was registered at ClinicalTrials.gov (NCT03793543), and the date of approval is 2 January 2019.

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

Data Availability Statement: Data are available to serious investigators upon reasonable request.

Conflicts of Interest: M.G.P. is co-inventor on a US patent covering ^{18}F -DCFPyL and as such is entitled to a portion of any licensing fees and royalties generated by this technology. R.A.W. has received advisory board compensation from Novartis/AAA and Bayer and speaker honoraria from Novartis/AAA and PentixaPharm. M.A.G. and S.P.R. are consultants to Progenics Pharmaceuticals, the licensee of ^{18}F -DCFPyL and a wholly owned subsidiary of Lantheus. T.G.P. and A.W. are employees of AIQ Solutions. The other authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PET	positron emission tomography
PSMA	prostate-specific membrane antigen
SUV	standardized uptake value
LOA	limits of agreement
ROI	region of interest

References

1. Schafer, E.J.; Laversanne, M.; Sung, H.; Soerjomataram, I.; Briganti, A.; Dahut, W.; Bray, F.; Jemal, A. Recent patterns and trends in global prostate cancer incidence and mortality: An update. *Eur. Urol.* **2025**, *87*, 302–313. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Raychaudhuri, R.; Lin, D.W.; Montgomery, R.B. Prostate cancer: A review. *JAMA* **2025**, *333*, 1433–1446. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Schaeffer, E.M.; Srinivas, S.; Adra, N.; An, Y.; Bitting, R.; Chapin, B.; Cheng, H.H.; D'Amico, A.V.; Desai, N.; Dorff, T. Prostate cancer, version 3.2024 featured updates to the NCCN guidelines. *JNCCN J. Natl. Compr. Canc Netw.* **2024**, *22*, 140–150. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Rowe, S.P.; Gorin, M.A.; Pomper, M.G. Imaging of prostate-specific membrane antigen with small-molecule PET radiotracers: From the bench to advanced clinical applications. *Annu. Rev. Med.* **2019**, *70*, 461–477. [\[CrossRef\]](#)
5. Sartor, O.; de Bono, J.; Chi, K.N.; Fizazi, K.; Herrmann, K.; Rahbar, K.; Tagawa, S.T.; Nordquist, L.T.; Vaishampayan, N.; El-Haddad, G.; et al. Lutetium-177-PSMA-617 for Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* **2021**, *385*, 1091–1103. [\[CrossRef\]](#)
6. Morris, M.J.; Rowe, S.P.; Gorin, M.A.; Saperstein, L.; Pouliot, F.; Josephson, D.; Wong, J.Y.; Pantel, A.R.; Cho, S.Y.; Gage, K.L. Diagnostic performance of ^{18}F -DCFPyL-PET/CT in men with biochemically recurrent prostate cancer: Results from the CONDOR phase III, multicenter study. *Clin. Cancer Res.* **2021**, *27*, 3674–3682. [\[CrossRef\]](#)
7. Farolfi, A.; Calderoni, L.; Mattana, F.; Mei, R.; Telo, S.; Fanti, S.; Castellucci, P. Current and Emerging Clinical Applications of PSMA PET Diagnostic Imaging for Prostate Cancer. *J. Nucl. Med.* **2021**, *62*, 596–604. [\[CrossRef\]](#)
8. Gafita, A.; Schroeder, J.A.; Ceci, F.; Oldan, J.D.; Khandani, A.H.; Lecouvet, F.E.; Solnes, L.B.; Rowe, S.P. Treatment Response Evaluation in Prostate Cancer Using PSMA PET/CT. *J. Nucl. Med.* **2025**, *66*, 995–1004. [\[CrossRef\]](#)
9. Smith, T.; Harper, M. Prostate-Specific Membrane Antigen (PSMA) PET-CT: Revolutionizing Staging, Restaging, and Treatment Response Assessment. *Ann. Urol. Oncol.* **2025**, *8*, 200–210. [\[CrossRef\]](#)
10. Sahakyan, K.; Li, X.; Lodge, M.A.; Werner, R.A.; Bundschuh, R.A.; Bundschuh, L.; Kulkarni, H.R.; Schuchardt, C.; Baum, R.P.; Pienta, K.J.; et al. Semiquantitative Parameters in PSMA-Targeted PET Imaging with [^{18}F]DCFPyL: Inpatient and Interpatient Variability of Normal Organ Uptake. *Mol. Imaging Biol.* **2020**, *22*, 181–189. [\[CrossRef\]](#)
11. Seifert, R.; Sandach, P.; Kersting, D.; Fendler, W.P.; Hadaschik, B.; Herrmann, K.; Sunderland, J.J.; Pollard, J.H. Repeatability of ^{68}Ga -PSMA-HBED-CC PET/CT-derived total molecular tumor volume. *J. Nucl. Med.* **2022**, *63*, 746–753. [\[CrossRef\]](#)

12. Werner, R.A.; Habacha, B.; Lutje, S.; Bundschuh, L.; Higuchi, T.; Hartrampf, P.; Serfling, S.E.; Derlin, T.; Lapa, C.; Buck, A.K.; et al. High SUVs Have More Robust Repeatability in Patients with Metastatic Prostate Cancer: Results from a Prospective Test-Retest Cohort Imaged with ^{18}F -DCFPyL. *Mol. Imaging* **2022**, *2022*, 7056983. [\[CrossRef\]](#)
13. Werner, R.A.; Habacha, B.; Lütje, S.; Bundschuh, L.; Kosmala, A.; Essler, M.; Derlin, T.; Higuchi, T.; Lapa, C.; Buck, A.K. Lack of repeatability of radiomic features derived from PET scans: Results from a ^{18}F -DCFPyL test–retest cohort. *Prostate* **2023**, *83*, 547–554. [\[CrossRef\]](#)
14. Trägårdh, E.; Ulén, J.; Enqvist, O.; Larsson, M.; Valind, K.; Minarik, D.; Edenbrandt, L. A fully automated AI-based method for tumour detection and quantification on [^{18}F] PSMA-1007 PET–CT images in prostate cancer. *EJNMMI Phys.* **2025**, *12*, 78. [\[CrossRef\]](#)
15. Jafari, E.; Zarei, A.; Dadgar, H.; Keshavarz, A.; Manafi-Farid, R.; Rostami, H.; Assadi, M. A convolutional neural network-based system for fully automatic segmentation of whole-body [^{68}Ga]Ga-PSMA PET images in prostate cancer. *Eur. J. Nucl. Med. Mol. Imaging* **2024**, *51*, 1476–1487. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Huang, B.; Yang, Q.; Li, X.; Wu, Y.; Liu, Z.; Pan, Z.; Zhong, S.; Song, S.; Zuo, C. Deep learning-based whole-body characterization of prostate cancer lesions on [^{68}Ga]Ga-PSMA-11 PET/CT in patients with post-prostatectomy recurrence. *Eur. J. Nucl. Med. Mol. Imaging* **2024**, *51*, 1173–1184. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Werner, R.A.; Lutje, S.; Habacha, B.; Bundschuh, L.; Higuchi, T.; Buck, A.K.; Kosmala, A.; Lapa, C.; Essler, M.; Lodge, M.A.; et al. Test-retest repeatability of organ uptake on PSMA-targeted ^{18}F -DCFPyL PET/CT in patients with prostate cancer. *Prostate* **2023**, *83*, 1186–1192. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Weisman, A.; Lokre, O.; Schott, B.; Fernandes, V.; Jeraj, R.; Perk, T.; Cho, S.; Perlman, S. Automated detection and quantification of neuroendocrine tumors on ^{68}Ga -DOTATATE PET/CT images using a U-net ensemble method. *J. Nucl. Med.* **2022**, *63*, 3215.
19. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 20–29 October 2017.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing And Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015.
21. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
22. Jaeger, P.F.; Kohl, S.A.; Bickelhaupt, S.; Isensee, F.; Kuder, T.A.; Schlemmer, H.-P.; Maier-Hein, K.H. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. In *Machine Learning for Health Workshop, Online, 11 December 2020*; PMLR: London, UK, 2020.
23. Huff, D.T.; Santoro-Fernandes, V.; Chen, S.; Chen, M.; Kashuk, C.; Weisman, A.J.; Jeraj, R.; Perk, T.G. Performance of an automated registration-based method for longitudinal lesion matching and comparison to inter-reader variability. *Phys. Med. Biol.* **2023**, *68*, 175031. [\[CrossRef\]](#)
24. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420. [\[CrossRef\]](#)
25. Bland, J.M.; Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *327*, 307–310. [\[CrossRef\]](#)
26. Lin, C.; Bradshaw, T.; Perk, T.; Harmon, S.; Eickhoff, J.; Jallow, N.; Choyke, P.L.; Dahut, W.L.; Larson, S.; Humm, J.L.; et al. Repeatability of Quantitative ^{18}F -NaF PET: A Multicenter Study. *J. Nucl. Med.* **2016**, *57*, 1872–1879. [\[CrossRef\]](#)
27. Zou, G. Confidence Interval Estimation for the Bland–Altman Limits of Agreement with Multiple Observations per Individual. In *Statistical Methods in Medical Research*; Wiley: Hoboken, NJ, USA, 2013; Volume 22, pp. 630–642.
28. Sun, R.; Lerousseau, M.; Briend-Diop, J.; Routier, E.; Roy, S.; Henry, T.; Ka, K.; Jiang, R.; Temar, N.; Carré, A. Radiomics to evaluate interlesion heterogeneity and to predict lesion response and patient outcomes using a validated signature of CD8 cells in advanced melanoma patients treated with anti-PD1 immunotherapy. *J. Immunother. Cancer.* **2022**, *10*, e004867. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Phillips, R.; Shi, W.Y.; Deek, M.; Radwan, N.; Lim, S.J.; Antonarakis, E.S.; Rowe, S.P.; Ross, A.E.; Gorin, M.A.; Deville, C.; et al. Outcomes of Observation vs Stereotactic Ablative Radiation for Oligometastatic Prostate Cancer: The ORIOLE Phase 2 Randomized Clinical Trial. *JAMA Oncol.* **2020**, *6*, 650–659. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Meikle, S.R.; Sossi, V.; Roncali, E.; Cherry, S.R.; Banati, R.; Mankoff, D.; Jones, T.; James, M.; Sutcliffe, J.; Ouyang, J.; et al. Quantitative PET in the 2020s: A roadmap. *Phys. Med. Biol.* **2021**, *66*, 06RM01. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Lodge, M.A.; Wahl, R.L. Practical PERCIST: A simplified guide to PET response criteria in solid tumors 1.0. *Radiology* **2016**, *280*, 576. [\[CrossRef\]](#)
32. Gafita, A.; Djaileb, L.; Rauscher, I.; Fendler, W.P.; Hadaschik, B.; Rowe, S.P.; Herrmann, K.; Calais, J.; Rettig, M.; Eiber, M. Response evaluation criteria in PSMA PET/CT (RECIP 1.0) in metastatic castration-resistant prostate cancer. *Radiology* **2023**, *308*, e222148. [\[CrossRef\]](#)

33. Shankar, L.K.; Huang, E.; Litierie, S.; Hoekstra, O.S.; Schwartz, L.; Collette, S.; Boellaard, R.; Bogaerts, J.; Seymour, L.; deVries, E.G. Meta-analysis of the test–retest repeatability of [^{18}F]-fluorodeoxyglucose standardized uptake values: Implications for assessment of tumor response. *Clin. Cancer Res.* **2023**, *29*, 143–153. [[CrossRef](#)]
34. Lin, C.; Harmon, S.; Bradshaw, T.; Eickhoff, J.; Perlman, S.; Liu, G.; Jeraj, R. Response-to-repeatability of quantitative imaging features for longitudinal response assessment. *Phys. Med. Biol.* **2019**, *64*, 025019. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.