BJR

https://doi.org/10.1259/bjr.20201174

Received: 29 September 2020 | Revised: 19 January 2021 | Accepted: 20 January 2021

Cite this article as:
Hearn N, Blazak J, Vivian P, Vignarajah D, Cahill K, Atwell D, et al. Prostate cancer GTV delineation with biparametric MRI and ^{68}Ga-PSMA-PET: comparison of expert contours and semi-automated methods. *Br J Radiol* 2021; **94**: 20201174.

FULL PAPER

Prostate cancer GTV delineation with biparametric MRI and ^{68}Ga-PSMA-PET: comparison of expert contours and semi-automated methods

[1,2,3]NATHAN HEARN, BSc/MBBS (Hons), [4]JOHN BLAZAK, BAppSc, MBBS, FRANZCR, FAANMS, [4]PHILIP VIVIAN, MBBS, FRANZCR, [1,2]DINESH VIGNARAJAH, MBBS, FRANZCR, [1]KATELYN CAHILL, BAppSc (MRT-RT), [1,2,3]DAISY ATWELL, BSc/MBBS, [3,5]JIM LAGOPOULOS, BSc, MBiomedE, PhD, FAINM and [1,2,3]MYO MIN, MBBS, GradDip (Biotech), PhD, FRANZCR

[1]Department of Radiation Oncology, Sunshine Coast University Hospital, Birtinya, Australia
[2]ICON Cancer Centre, Maroochydore, Australia
[3]University of the Sunshine Coast, Sippy Downs, Australia
[4]Department of Medical Imaging, Sunshine Coast University Hospital, Birtinya, Australia
[5]Sunshine Coast Mind and Neuroscience – Thompson Institute, University of the Sunshine Coast, Birtinya, Australia

Address correspondence to: Dr Nathan Hearn
E-mail: *nathan.hearn@research.usc.edu.au*

Objective: The optimal method for delineation of dominant intraprostatic lesions (DIL) for targeted radiotherapy dose escalation is unclear. This study evaluated interobserver and intermodality variability of delineations on biparametric MRI (bpMRI), consisting of T_2 weighted (T_2W) and diffusion-weighted (DWI) sequences, and ^{68}Ga-PSMA-PET/CT; and compared manually delineated GTV contours with semi-automated segmentations based on quantitative thresholding of intraprostatic apparent diffusion coefficient (ADC) and standardised uptake values (SUV).

Methods: 16 patients who had bpMRI and PSMA-PET scanning performed prior to any treatment were eligible for inclusion. Four observers (two radiation oncologists, two radiologists) manually delineated the DIL on: (1) bpMRI (GTV_{MRI}), (2) PSMA-PET (GTV_{PSMA}) and (3) co-registered bpMRI/PSMA-PET (GTV_{Fused}) in separate sittings. Interobserver, intermodality and semi-automated comparisons were evaluated against consensus Simultaneous Truth and Performance Level Estimation (STAPLE) volumes, created from the relevant manual delineations of all observers with equal weighting. Comparisons included the Dice Similarity Coefficient (DSC), mean distance to agreement (MDA) and other metrics.

Results: Interobserver agreement was significantly higher ($p < 0.05$) for GTV_{PSMA} (DSC: 0.822, MDA: 1.12 mm) and GTV_{Fused} (DSC: 0.787, MDA: 1.34 mm) than for GTV_{MRI} (DSC: 0.705, MDA 2.44 mm). Intermodality agreement between GTV_{MRI} and GTV_{PSMA} was low (DSC: 0.440, MDA: 4.64 mm). Agreement between semi-automated volumes and consensus GTV was low for MRI (DSC: 0.370, MDA: 8.16 mm) and significantly higher for PSMA-PET (0.571, MDA: 4.45 mm, $p < 0.05$).

Conclusion: ^{68}Ga-PSMA-PET appears to improve interobserver consistency of DIL localisation *vs* bpMRI and may be more viable for simple quantitative delineation approaches; however, more sophisticated approaches to semi-automatic delineation factoring for patient- and disease-related heterogeneity are likely required.

Advances in knowledge: This is the first study to evaluate the interobserver variability of prostate GTV delineations with co-registered bpMRI and ^{68}Ga-PSMA-PET.

BACKGROUND

Radiotherapy of primary prostate cancer in conjunction with androgen deprivation therapy (ADT) is a well-established treatment option for the management of intermediate- and high-risk disease. Biochemical disease control can be improved with radiotherapy dose escalation, but maximal dose is limited due to toxicity.[1] Alternatively, the dominant intraprostatic lesion (DIL) may be focally dose-escalated with image-guided radiotherapy techniques, optimising the therapeutic ratio. This relies on accurate and reproducible definition of the DIL.

Functional imaging techniques have been increasingly incorporated into radiotherapy planning for this purpose:[2] specifically, dose escalation trials such as the FLAME and hypo-FLAME trials have used multiparametric MRI

(mpMRI) for lesion definition and graduated dose prescription.[3–5] However, significant variability between clinicians and institutions employing multimodal protocols has been described.[6] Availability of full mpMRI workup is not universal, and it is unclear whether biparametric MRI (bpMRI) examinations (consisting of T_2W and DWI series alone) provide sufficient characterisation of benign *vs* malignant tissue.[7]

Recently, ^{68}Ga-PSMA-PET has also demonstrated high sensitivity for the detection of both intraprostatic and metastatic lesions and has been increasingly employed in the work-up of high-risk disease.[8] PSMA-PET may therefore represent a growing resource for functional imaging-guided radiotherapy. The combination of mpMRI and PSMA-PET appears to be more accurate for lesion localisation than either alone[9] and may provide additional biological characterisation of malignant tissue for 'dose painting' or tumour risk scoring tools.

Few studies have addressed whether bpMRI alone or in combination with PSMA-PET could be used for reproducible lesion localisation for radiotherapy planning purposes. Semi-automated approaches to segmentation incorporating quantitative data from these modalities also require further exploration. This study was therefore undertaken to evaluate the interobserver and intermodality variability of DIL delineations undertaken on bpMRI, PSMA-PET and fused imaging sessions; and to compare manual DIL delineations to semi-automated approaches based on quantitative thresholding of intraprostatic SUV and ADC values.

METHODS

Study population

This study was approved by a local Human Research Ethics Committee. A list of all prostate cancer patients treated at our institution (Sunshine Coast University Hospital, Birtinya, Australia) within the prior 2 years was compiled through local database search. Patient were eligible for inclusion if they had suitable MRI and PSMA-PET acquired prior to undertaking any therapy (radiotherapy, ADT or prostatectomy). Exclusions were performed for cases of inadequate imaging quality (*e.g.* due to hip prosthesis), missing imaging sequences or time gap of more than 12 months between imaging acquisitions. Tumour laterality as assessed on transperineal or transrectal prostate biopsy was recorded for each patient for later comparison with contour localisation.

Image acquisition

MRI was acquired on a 3 T Siemens Skyra MRI and consisted of at least an axial T_2 weighted turbo spin echo (TSE) (TR/TE: 3200/108 ms, in-plane resolution: 0.56 mm; matrix size: 370 × 320; slice thickness: 3 mm; slices: 23–28) and an echoplanar DWI acquisition (TR/TE: 4300/60 ms, in-plane resolution: 1.93 mm; matrix size: 114 × 114; slice thickness: 3 mm, slices: 20–27; b values: 50, 400, 800). ^{68}Ga-PSMA-PET images were acquired from vertex to knees 45 min after administration of radiotracer (mean ± SD:159 ± 30 MBq) in conjunction with low-dose CT for attenuation correction on either a Siemens Biograph mCT Flow PET/CT or General Electric Discovery 710 PET/CT scanner.

All imaging sets were imported into MIM Maestro® (MIM Software Inc, Cleveland, OH) for contouring. Linked T_2 weighted and ADC sequences were checked for accurate registration, and minor manual adjustments made by the study radiation oncologists if required. The PSMA-PET/CT fusion was then rigidly registered to the bpMRI for each patient semi-automatically using inbuilt MIM registration based on mutual information. The study radiation oncologists checked all image fusions and performed final manual adjustment if required. The same fusion alignment was then used across all contouring sessions for all observers. Standardised windowing of SUV 0–5 was used across all PSMA-PET studies.

Manual GTV delineation

Four observers participated in the study: two radiation oncologists (6 and 4 years of specialist experience, including clinical and research study delineation tasks) and two radiologists (4 years each of specialist radiodiagnostic experience, but no delineation experience). Observers completed a practice set of contours on one patient's imaging, excluded from the analysis, to become familiar with the task and software.

During the contouring task, observers were blinded to patient information and completed contouring independently of each other in accordance with a standardised instruction sheet. For each patient, observers delineated GTV in three separate sessions: (1) MRI only (GTV_{MRI}), using fused axial T_2W and ADC images, (2) PSMA-PET/CT only (GTV_{PSMA}), and (3) with a co-registered MRI/PSMA-PET fusion (GTV_{Fused}), specifically consisting of T_2W/ADC and T_2W/PSMA-PET overlays.

Each observer completed a questionnaire scoring image quality, confidence in contours and fusion quality using a 4-point scale. Additionally, observers were asked to list their preferred delineation session (MRI, PSMA or Fused) for each patient and record the time taken for each contour (see Supplementary Appendix A: Contouring questionnaire). Mean and standard deviation of these scores were descriptively reported.

Semi-automated delineation

One of the study radiation oncologists manually delineated the whole-prostate volume (WPV) on each axial slice of T_2W imaging for every patient, including gross instances of extraprostatic disease extension. WPVs were then propagated to co-registered ADC and PSMA-PET images and subregions of interest were defined through quantitative thresholding, based on the findings of previous histopathological and other studies. Specifically, subvolumes defined by ADC < 1000 mm^2/s x 10^{-6} (auto-GTV_{MRI})[10–17] and SUV >30%SUV_{max} (auto-GTV_{PSMA}),[18,19] and the Boolean unions and intersections of the prior volumes (auto-GTV_{uni} and auto-GTV_{int}) within the WPV contour were created. Non-contiguous volumes <0.5 ml created with this process were automatically removed prior to analysis.

Contour comparisons and statistical analysis

Contours were compared by volume, overlap and boundary agreement using well-established metrics.[20] The reference contour for each modality was taken to be a consensus delineation created

with the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm[21] with equal weighting for all observers' manual delineations. Volume agreement was evaluated with the intraclass correlation coefficient (two-way mixed method, absolute agreement, single measures) and with Bland–Altman plots with 95% limits of agreement. Contour overlap was quantified with the Dice Similarity Coefficient (DSC) and Jaccard Index (JI). Boundary agreement was measured with the maximum Hausdorff Distance (HDmax) and Mean Distance to Agreement (MDA).

To assess interobserver variability, these measures were calculated for each observer-STAPLE pair and averaged. Intermodality variability was evaluated using the same metrics, comparing the consensus STAPLE contour of one session to another. Finally, STAPLE contours for each session were compared to their corresponding semi-automated segmentation. Differences were statistically evaluated in the R Statistical Environment (R Core Team, v. 3.6.1) with the Wilcoxon signed rank test with repeated measures ($p < 0.05$, two-tailed).

RESULTS

Patient characteristics

Initial database search identified 171 patients diagnosed with prostate cancer within the specified time frame, with a total of 42 patients undergoing both MRI and PSMA-PET examinations. After exclusions (Figure 1), a resultant 16 patients were suitable for inclusion in the study, 4 of whom underwent radical prostatectomy and the remainder radiotherapy and/or ADT (Table 1). This also included one patient who underwent both imaging examinations with findings consistent with prostate cancer, associated with a serially rising PSA (17 ng ml^{-1} at time of imaging), but who eventually declined biopsy and treatment.

Manual GTV delineation – interobserver agreement

Examples of GTV delineation on MRI, PSMA-PET and fused imaging sessions are shown in Figure 2. Volumes for each observer/modality combination and STAPLE consensus contours are shown in Figure 3. Contour volumes and interobserver variability statistics are summarised in Figure 4 and Table 2.

Image quality was generally satisfactory for all modalities but was rated slightly higher for PSMA-PET (3.35, SD: 0.81) than for MRI (2.79, SD: 0.80) or fused image sets (2.88, SD: 0.91). Time to complete contours did not greatly differ between sessions (MRI mean 4.4 min, PSMA-PET mean 3.8 min, fusion mean 4.2 min). Observer-reported confidence was highest in PSMA-PET delineations (3.35, SD: 0.81) followed by fused (2.85, SD: 0.92) and MRI sessions (2.60, SD: 0.89). Observer preference for

Figure 1. Study eligibility assessment. PET, positron emission tomography; PSMA, prostate specific membrane antigen.

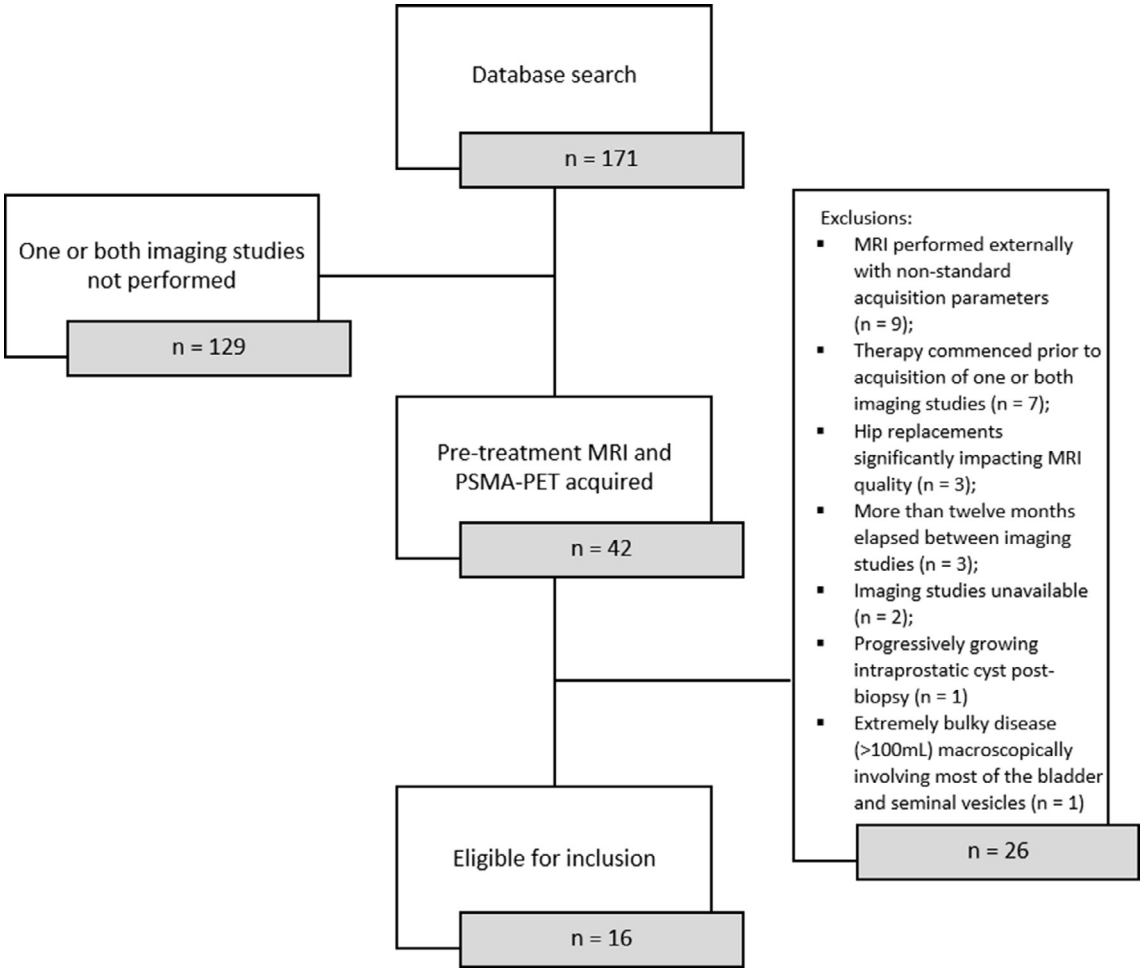


Table 1. Patient and tumour characteristics

Patients	16
Age (median, range)	67 (53–82)
T stage, n (%)	
T2	7 (43.8)
T3	8 (50.0)
T4	1 (6.3)
N stage, n (%)	
N0	10 (62.5)
N1	6 (37.5)
Gleason grading, n (%)	
3 + 3 = 6	0
3 + 4 = 7	3 (18.8)
4 + 3 = 7	5 (31.3)
4 + 4 = 8	1 (6.3)
4 + 5 = 9	5 (31.3)
5 + 4 = 9	0
5 + 5 = 10	1 (6.3)
Declined biopsy	1 (6.3)
Localisation	
Anterior fibromuscular stroma	0
Central zone	1 (6.3)
Peripheral zone	11 (68.8)
Transition zone	7 (43.8)
Satellite lesions	3 (18.8)
Treatment	
Radical prostatectomy	4 (25.0)
Androgen deprivation therapy ± radiotherapy	11 (68.8)
Declined treatment	1 (6.3)

delineation modality was most commonly the fused image set (46%) followed by PSMA-PET (33%) and MRI (21%).

There were no significant differences between GTV sizes delineated on MRI, PSMA or Fused imaging sets ($p > 0.05$). Interobserver volumetric agreement was high across all modalities, with GTV_{MRI}, GTV_{PSMA} and GTV_{Fused} ICC of 0.78 (95% CI: 0.64–0.89), 0.80 (95% CI: 0.66–0.90) and 0.90 (95% CI: 0.83–0.95) respectively. Bland–Altman plots did not demonstrate clear directional bias for observer–STAPLE comparisons (Figure 4). Limits of agreement for GTV_{MRI}, GTV_{PSMA} and GTV_{Fused} volumes were −15.0/+14.4 ml, −25.7/+24.5 ml and −18.8/+15.6 ml respectively.

GTV_{PSMA} contours demonstrated significantly higher interobserver spatial overlap agreement (DSC: 0.822, JI: 0.715) than GTV_{MRI} (DSC: 0.705, JI: 0.587) or GTV_{Fused} (DSC: 0.787, JI: 0.667) contours ($p < 0.05$ all comparisons). Overlap agreement was also significantly higher in fused session contours than with MRI alone (DSC: $p = 0.016$, JI: $p = 0.018$).

No significant difference ($p > 0.05$) in HDmax was found between GTV_{MRI} (8.43 mm), GTV_{PSMA} (6.78 mm) and GTV_{Fused} (7.03 mm) delineations. MDA was significantly lower for GTV_{PSMA} (1.12 mm) delineations than GTV_{MRI} (2.44 mm, $p = 0.025$) but not *vs* GTV_{Fused} (1.34 mm, $p = 0.075$).

Manual GTV delineation – intermodality agreement

Intermodality agreement between GTV overlap and boundary on MRI and PSMA was moderate at best (DSC: 0.440, JI: 0.312, HDmax: 14.87 mm, MDA: 4.64 mm). GTV_{Fused} contours were more closely aligned with GTV_{PSMA} than GTV_{MRI} (Table 2). In considering lesion localisation *vs* histopathological biopsy, GTV was correctly lateralised by observers in 13/16 cases on MRI, in 14/16 cases on PSMA-PET and in 15/16 cases with fused imaging. In Bland–Altman analysis, limits of agreement were −15.6/+8.9 ml (Figure 4F).

Semi-automated GTV delineation – agreement with consensus volumes

Examples of semi-automated GTV delineation according to intraprostatic ADC <1000 mm^2/s (auto-GTV_{MRI}), SUV >30%SUV_{max} (auto-GTV_{PSMA}) and the Boolean unions and intersections of the two (Auto-GTV_{uni} and Auto-GTV_{int}) are shown in Figure 5 and comparisons with corresponding consensus contours are summarised in Figure 6 and Table 2.

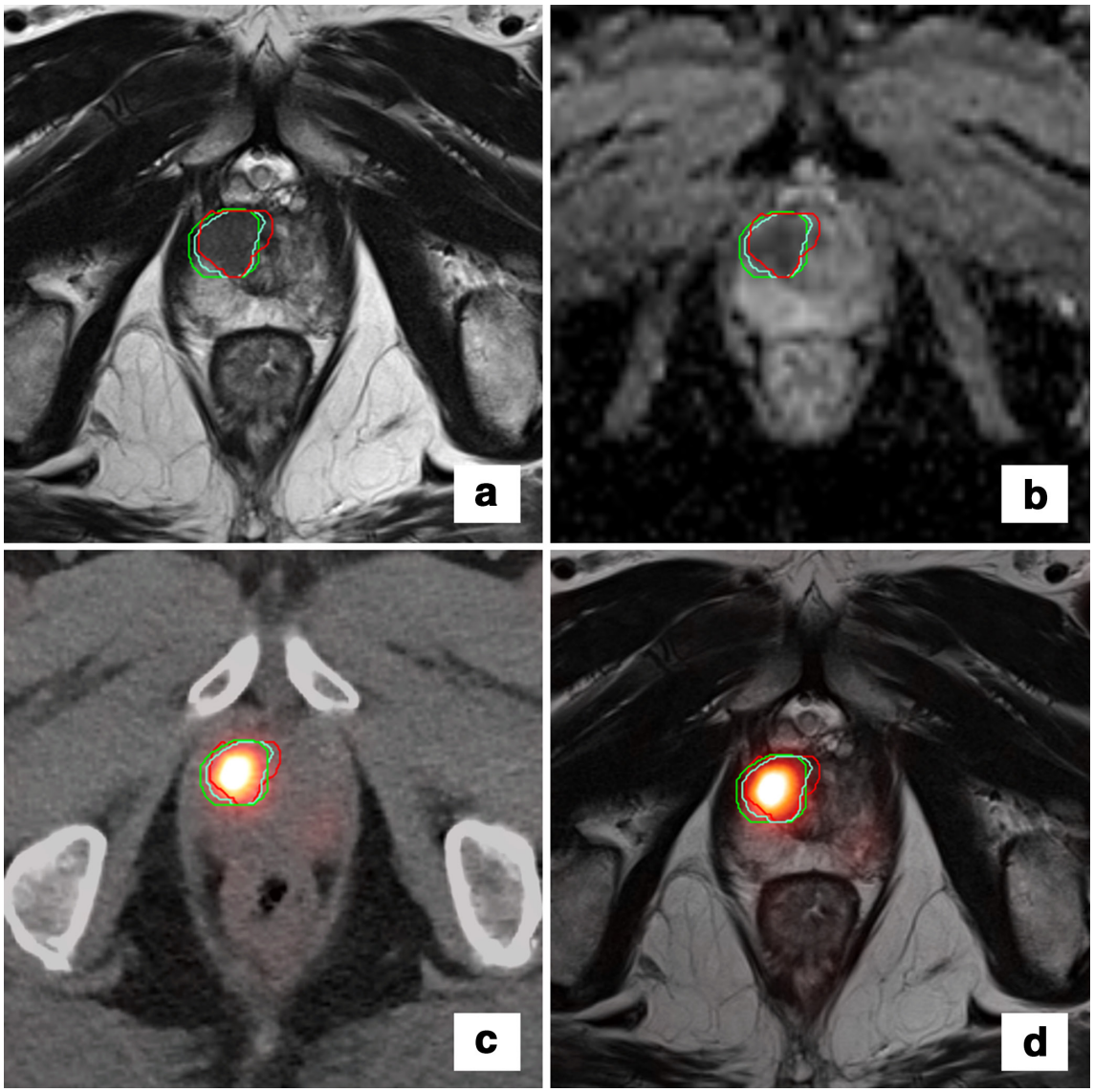
Auto-GTV_{MRI} contours demonstrated poor spatial overlap and boundary agreement with consensus manual delineations (DSC 0.370, JI 0.262, HDmax: 24.61 mm, MDA: 8.16 mm). Auto-GTV_{PMSA} contours demonstrated significantly higher agreement to their corresponding manual delineations ($p < 0.05$), with moderate agreement (DSC: 0.571, JI: 0.435, HDmax: 14.99 mm, MDA: 4.45 mm). Auto-GTV_{uni} volumes demonstrated closer overlap, but worse boundary agreement with consensus GTV_{Fused} session delineations (DSC: 0.489, JI: 0.360, HDmax: 23.79 mm, MDA: 7.30 mm) compared to Auto-GTV_{int} (DSC: 0.384, JI: 0.265, HDmax: 16.89 mm, MDA: 5.93 mm).

Cases were reviewed to establish the main causes of poor agreement with manual delineation. Most commonly, ADC- and SUV-based thresholding failed to produce coherent volumes in smaller and less-avid lesions given the lack of contrast with normal tissue (Figure 5D–F). Overinclusion of the anterior fibromuscular stroma and hyperplastic nodules in the transition zone was also seen. PSMA-PET segmentation was less reliable if the DIL was proximal to the bladder, due to intravesicular contrast accumulation and resolution constraints. Finally, misregistration between MRI and PSMA-PET was noted for two cases due to unequal bladder filling and rectal luminal variation.

DISCUSSION

To the best of our knowledge, this is the first study to consider the interobserver variability of intraprostatic lesion delineations with the combination of bpMRI and ^{68}Ga-PSMA-PET. The results of this study demonstrate moderate interobserver

Figure 2. Axial image examples of GTV delineation on bpMRI (A-B), PET/CT (C) and Fused bpMRI/PET (D) sessions. Red: STAPLE GTV_{MRI}; green: STAPLE GTV_{PSMA}; cyan: GTV_{Fused}. GTV, gross tumour volume; PET, positron emission tomography; STAPLE, Simultaneous Truth and Performance Level Estimation.

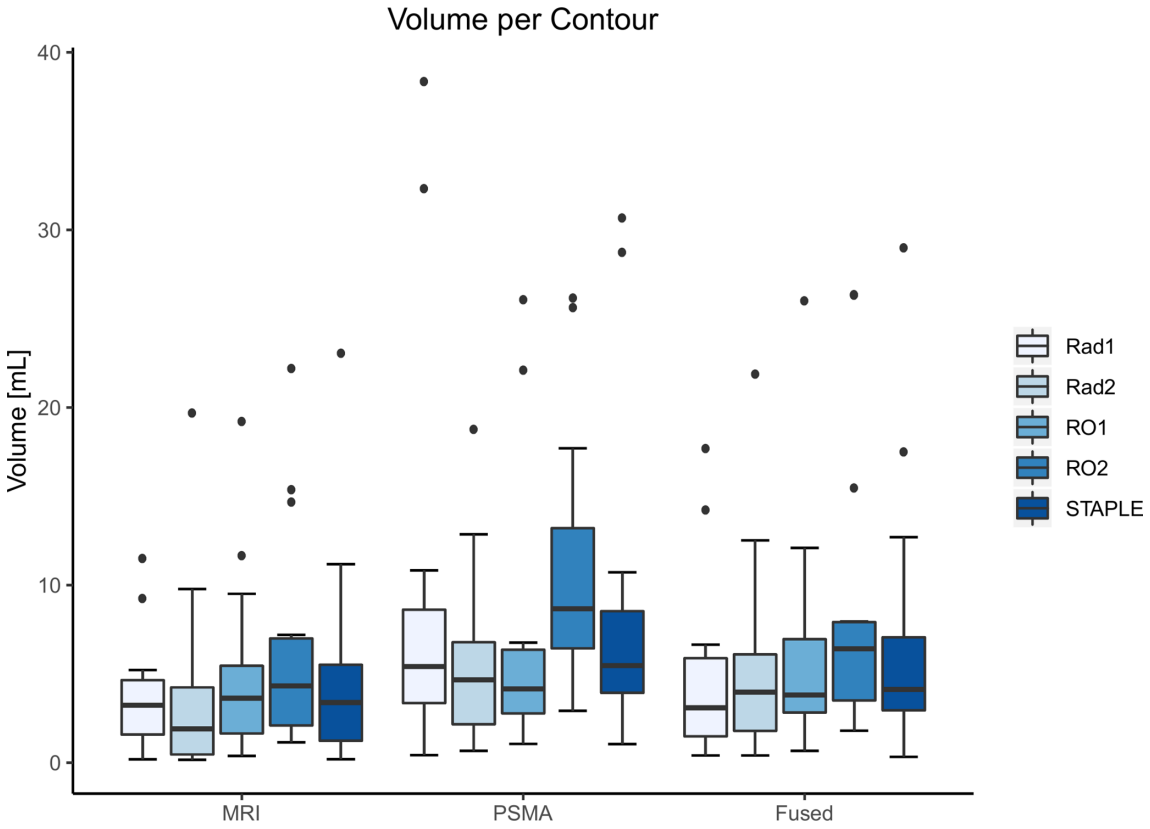


agreement of DIL localisation with bpMRI and suggest semi-automated lesion delineation with a simple ADC thresholding approach is not sufficient for accurate lesion localisation, consistent with previous findings of mpMRI interpretation and lesion localisation. Contours delineated on PSMA-PET alone or in conjunction with fused bpMRI had greater interobserver consistency and showed closer alignment between manual and semi-automated volumes.

While we could not undertake direct comparison of bpMRI and mpMRI, interobserver variability in bpMRI delineations in the present study appeared to be similar to that of mpMRI delineations described in the literature, such as those of Piert et al.[22] comparing contours between two observers (DSC: 0.41, JI: 0.27, HD: 8.56 mm, MDA: 2.43 mm),[22] and other similar analyses.[23,24] Comparatively, interobserver agreement by overlap measures and MDA appeared to be significantly greater in GTV_{PSMA} and GTV_{Fused} sessions *vs* GTV_{MRI}. The degree of contrast between cancerous and benign tissue is likely the main determinant of these improvements: *i.e.* greater lesion conspicuity on PSMA-PET may have improved overall consistency between observers. Meanwhile, similar HDmax between sessions may reflect an inherent variability between clinicians in defining the maximal extent of tumour invasion on both MRI and PSMA-PET.

Recent studies have considered the impact of technical considerations such as image scaling on interobserver variability in contouring tasks: Zamboglou et al.[25] described significantly greater agreement when uniform SUV windowing of 0–5 (DSC 0.80) was employed (as was used here) compared to non-standardised scaling methods (DSC 0.56).[25] Draulans et al.[26] found alternatively that windowing of 0–41% SUV_{max} optimised the accuracy of contours with reference to histopathology.[26] Similar considerations may apply to bpMRI, where choice of

Figure 3. Boxplot of contour volumes per modality and per observer. Rad: radiologist; RO: radiation oncologist; STAPLE: consensus volume via STAPLE algorithm with equal weighting for all observers. STAPLE, Simultaneous Truth and Performance Level Estimation.



b-values or the incorporation of simulated high b-value DWI could improve consistency between observers on bpMRI.[27] Ultimate optimisation of image scaling and display is likely dependent on specific scanner, image acquisition and patient characteristics, and should be further investigated.

While lateralisation accuracy with reference to biopsy specimens was similar between modalities in this study, the small subject pool and lack of definitive histopathological assessments limit the conclusions that can be drawn with respect to intermodality agreement. Zamboglou et al previously found agreement rates of 47% between laterality on mpMRI, PSMA-PET and histology in a retrospective study of 22 patients;[28] and, in a later study, found clinician-defined contours on PSMA-PET covered a higher proportion of malignant areas than mpMRI.[9] It is unclear whether the greater agreement between GTV_{PSMA}/GTV_{Fused} versus GTV_{MRI}/GTV_{Fused} in the present study is due to underlying sensitivity differences between bpMRI and PSMA-PET, or reflective of observer preference for PSMA-PET and more obvious contrast with diseased tissue. However, in light of existing histopathological data, the similar time to task completion between contouring sessions, and the associated improvements in interobserver variability, it is reasonable to suggest that PSMA-PET should be utilised to assist in manual DIL delineations where available.

Quantitative thresholding methods based on intraprostatic ADC and $\%SUV_{max}$ have been investigated as a means to standardise functional imaging-based contouring tasks. While the threshold-based auto-GTV_{PSMA} contours marginally outperformed auto-GTV_{MRI} with reference to manual consensus delineations, we found generally moderate overlap and poor boundary concordance with both methods. Of note, the large MDA (8.16 mm) and HDmax (24.61 mm) of auto-GTV_{MRI} contours indicate significant overinclusion of tissue classed as non-malignant by expert observer consensus, even after cleaning of minor non-contiguous volumes. In some instances, low inherent contrast between malignant tissue of low avidity against high background uptake produced no useful volume (Figure 5F). We subsequently could not establish any benefit to the combined use with Boolean union/intersection of semi-automated volumes, despite previous findings of improved histological coverage with mpMRI/PSMA-PET volume unions.[9]

However, a wide range of cut-off values for this purpose is present in the literature: ADC values from approximately 700–1300 $mm^2/s \times 10^{-6}$ [10–17] and $\%SUV_{max}$ values from 20–50%,[18,19,25,26,29] have been variably suggested. Similarly to the issue of optimising image scaling for manual delineations, experimental factors such as acquisition parameters – b-values in the case of DWI, and other reconstruction factors in PET[30] – significantly impact the accuracy of derived thresholds. Specific to ^{68}Ga-PSMA, intravesicular accumulation of radioligand may obscure lesions at the prostate base and complicate approaches based on $\%SUV_{max}$. Disease-related factors, including both cellular density[31] and

Figure 4. Interobserver (a–e) and intermodality variability (f) of GTV contours completed on MRI, PSMA and Fused imaging sets. Boxplots of (a) DSC and (b) JI; Bland–Altman plots with mean and 95% limits of agreement for observer–STAPLE comparisons on (c) MRI, (d) PSMA-PET, and (e) Fused contouring tasks; and (f) intermodality comparison between MRI and PSMA contours for each observer. DSC, Dice Similarity Coefficient; GTV, gross tumour volume; JI, Jaccard Index; PET, positron emission tomography; PSMA, prostate specific membrane antigen; STAPLE, Simultaneous Truth and Performance Level Estimation.

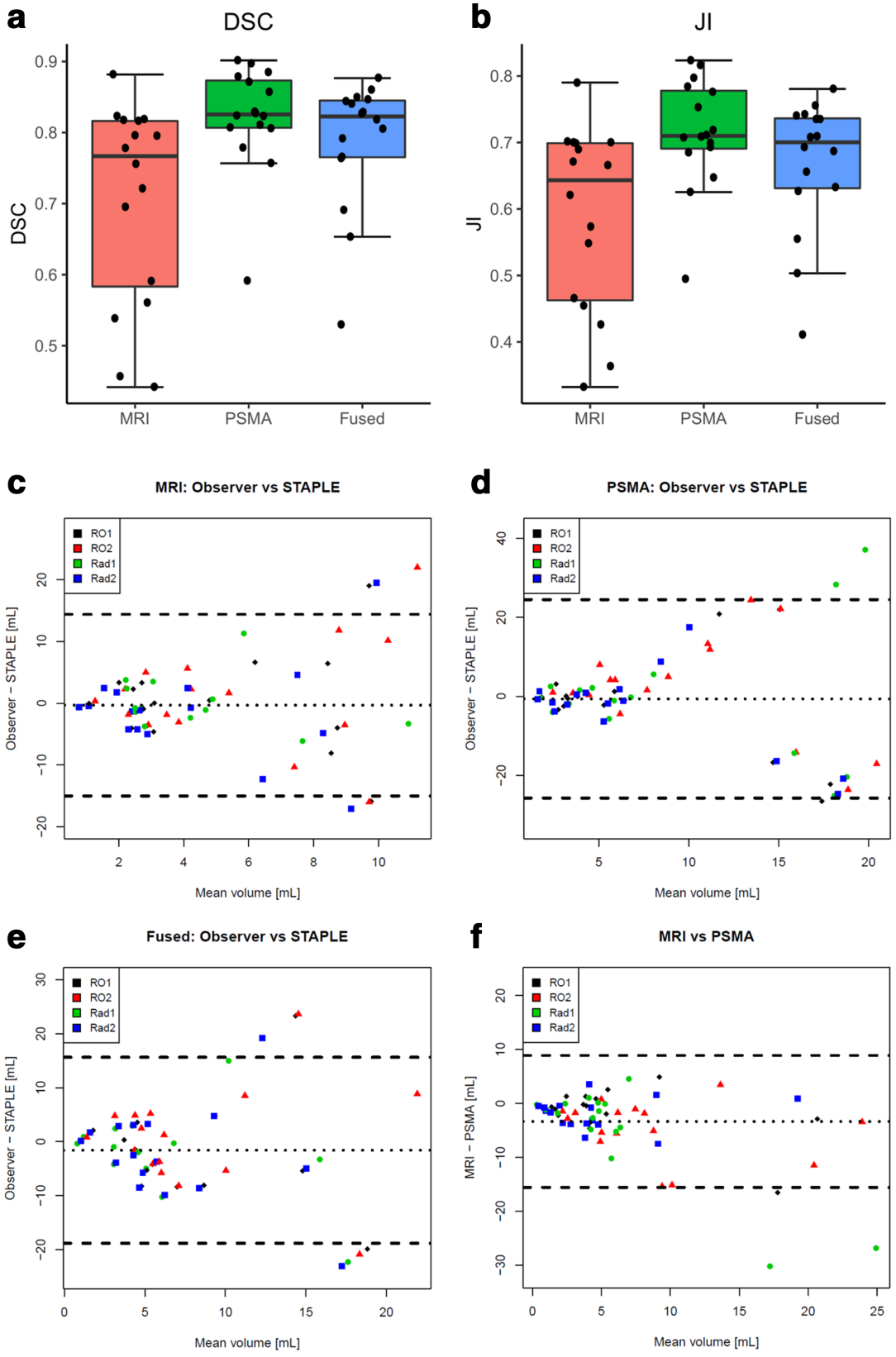


Table 2. Summary interobserver and intermodality variability statistics of manual GTV delineations undertaken on MRI, PSMA and fused imaging sets; and comparisons of STAPLE consensus delineations with semi-automated delineations based on ADC and %SUV_{max}

Contour	Volume (mL)	DSC	JI	HDmax (mm)	MDA (mm)
Manual delineation **interobserver** variability (observer *vs* STAPLE comparisons)					
MRI	4.90 (2.52–7.28)	0.705 (0.636–0.775)	0.587 (0.519–0.656)	8.69 (6.55–10.83)	2.44 (1.17–3.71)
PSMA	8.58 (3.81–13.35)	0.822 (0.785–0.858)*	0.715 (0.675–0.755)*	6.60 (4.93–8.27)	1.12 (0.68–1.56)*
Fused	7.79 (4.26–11.33)	0.787 (0.742–0.832)*†	0.667 (0.617–0.716)*†	6.71 (5.34–8.07)	1.34 (0.66–2.02)†
Manual delineation **intermodality** variability (STAPLE *vs* STAPLE comparisons)					
MRI *vs* PSMA	-	0.440 (0.318–0.562)	0.312 (0.234–0.410)	14.87 (10.75–18.99)	4.64 (2.60–6.68)
MRI *vs* Fused	-	0.605 (0.486–0.724)	0.470 (0.359–0.581)	11.31 (7.32–15.30)	3.01 (1.19–4.83)
PSMA *vs* Fused	-	0.714 (0.606–0.822)	0.588 (0.489–0.687)	7.40 (3.94–10.86)	1.93 (0.42–3.44)
Semi-automated delineations (pairwise automated *vs* manual STAPLE comparisons)					
Auto-GTV_{MRI} (ADC <1000)	5.37 (3.01–7.72)	0.370 (0.236–0.504)	0.262 (0.154–0.370)	24.61 (18.78–30.43)	8.16 (4.44–11.88)
Auto-GTV_{PSMA} (SUV >30%SUV_{max})	15.33 (8.20–22.47)	0.571 (0.450–0.692)*	0.435 (0.327–0.543)*	14.99 (9.37–20.62)*	4.45 (1.90–7.00)*
Auto-GTV_{uni}	17.61 (10.77–24.44)	0.489 (0.358–0.620)	0.360 (0.252–0.468)	23.79 (17.64–29.93)	7.30 (4.06–10.54)
Auto-GTV_{int}	3.22 (1.52–4.92)	0.384 (0.267–0.502)	0.265 (0.168–0.363)	16.89 (11.20–22.57)	5.93 (2.61–9.25)

ADC, apparent diffusion coefficient; GTV, gross tumour volume; PSMA, prostate specific membrane antige; STAPLE, Simultaneous Truth and Performance Level Estimation; SUV, standardised uptake value.
Data displayed as mean (95% CI). *Significant difference compared to MRI ($p < 0.05$); †Significant difference compared to PSMA ($p < 0.05$).

Figure 5. Examples of favourable (a–c) and unfavourable (d–f) cases of semi-automated delineations and corresponding STAPLE consensus volumes. (a, b), (d, e) axial views of STAPLE GTV_{MRI} (red) and Auto-GTV_{MRI} (shaded green); (c) axial and (f) sagittal views of STAPLE GTV_{PSMA} (green) and Auto-GTV_{PSMA} (shaded cyan). GTV, gross tumour volume; PSMA, prostate specific membrane antigen; STAPLE, Simultaneous Truth and Performance Level Estimation.

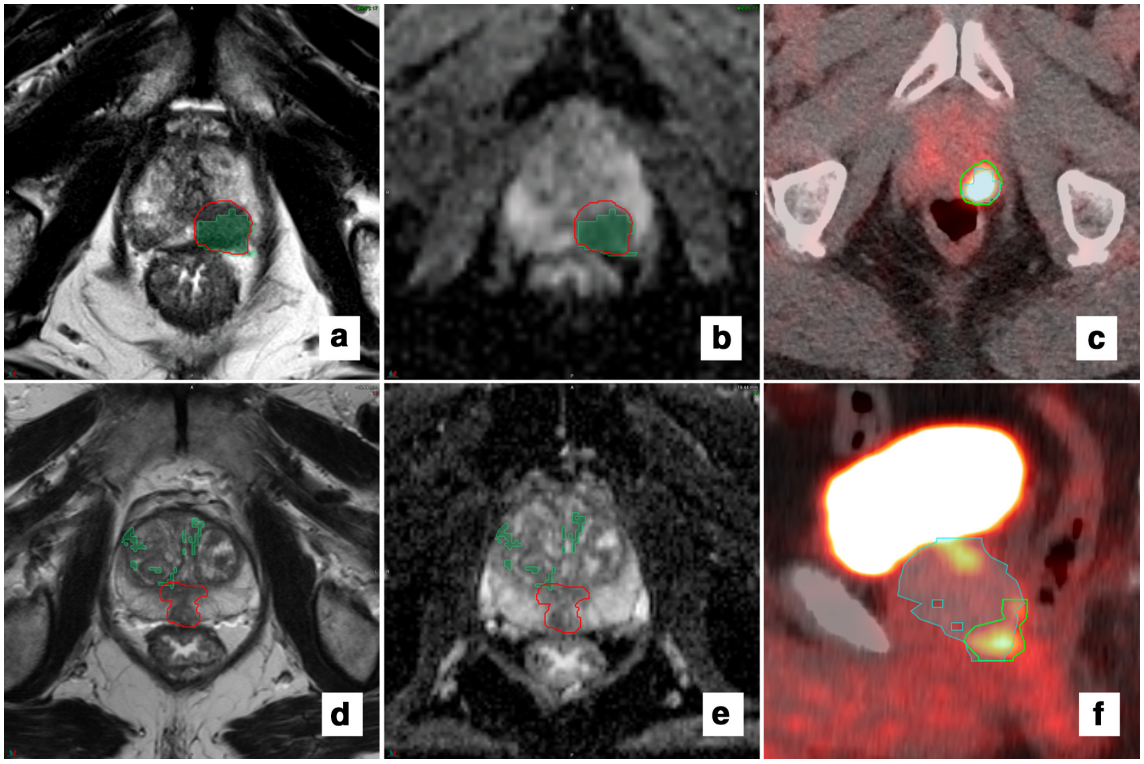
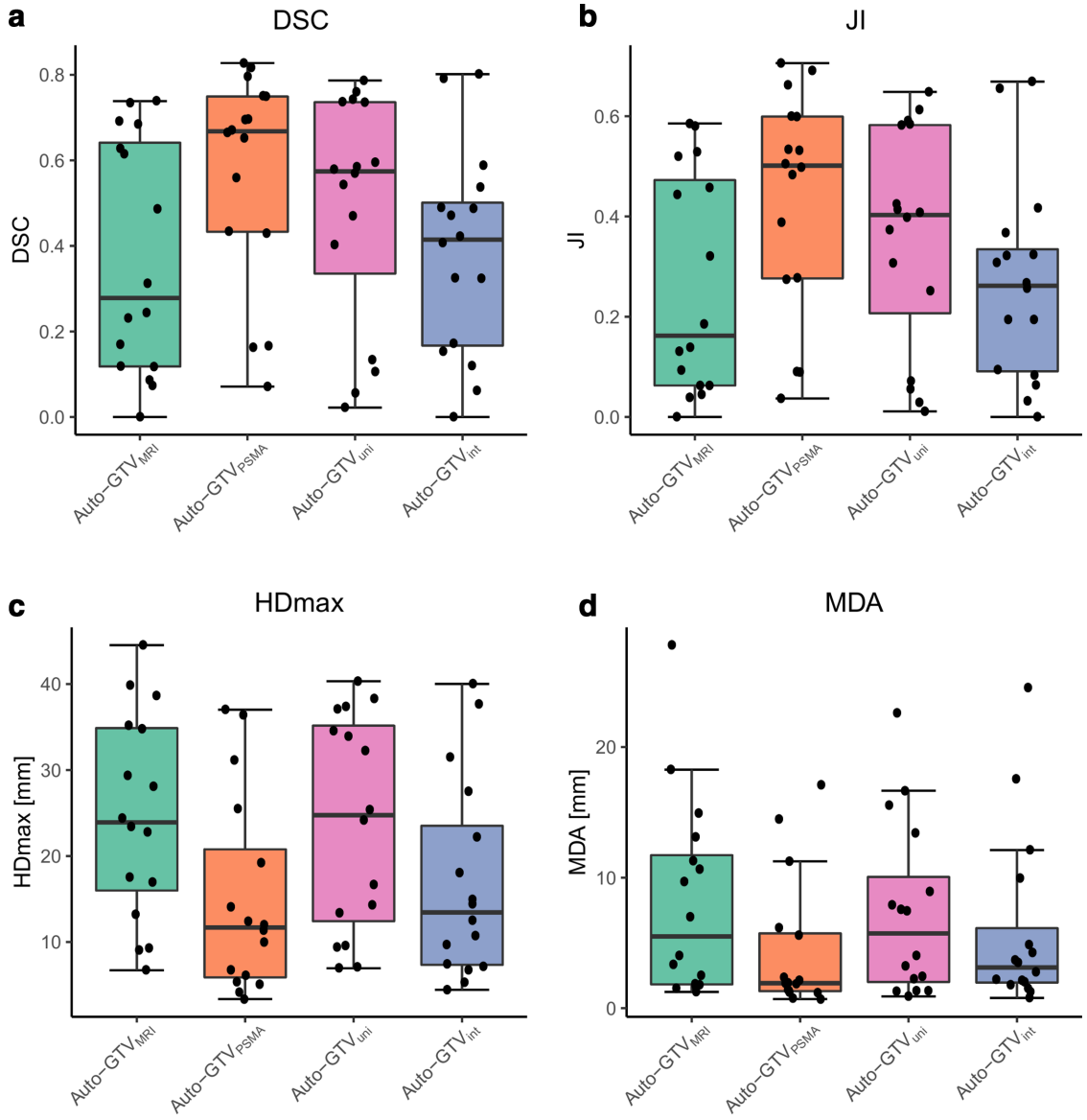


Figure 6. Summary boxplot of semi-automated contours generated through whole prostate volume thresholding according to ADC (Auto-GTV_{MRI}), %SUV_{max} (Auto-GTV_{PSMA}) and their Boolean union and intersections (Auto-GTV_{uni}, Auto-GTV_{int}). (a) DSC, (b) JI, (c) HDmax and (d) MDA calculated against the corresponding manually delineated STAPLE volume (GTV_{Fused} used for both union and intersection comparisons). GTV, gross tumour volume; JI, Jaccard Index; PET, positron emission tomography; PSMA, prostate specific membrane antigen; STAPLE, Simultaneous Truth and Performance Level Estimation.



Gleason score[32] affects tumour ADC measurements and SUV_{max},[33] which may impact the generalisability of uniform thresholding. Finally, inherent intra- and intersubject variability of both benign and cancerous tissue impairs direct correlation with functional imaging parameters such as ADC.[7]

Therefore, while the threshold values chosen in the present study could ostensibly be optimised *ad hoc*, a more sophisticated approach integrating the above factors is likely required. Tschudi et al recently described a differential risk scoring tool utilising separate graduated thresholds for transition and peripheral zone lesions.[34] Other authors have used patient-specific normalised values, relative to background prostate heterogeneity.[12] Further validation of these methods is required, as fundamental limitations related to voxel size, susceptibility artefact and benign mimics of disease may still remain.

The main strength of this study is the inclusion of four blinded observers comprised of both radiology and radiation oncology specialists. Limitations pertain to the source data available, including small sample size. While cases of MRI artefact (such as with hip prostheses) were excluded and MRI quality was generally judged to be of satisfactory quality, registration errors between T_2W and DWI sequences may impact the accuracy of intraprostatic thresholding based on ADC. PSMA-PET images were acquired from two different scanners, possibly increasing variability despite the use of patient-specific %SUV_{max} cut-off values. Time lapse between scans was also variable and may have

contributed to registration errors on fused images, particularly in cases with inconsistent bladder filling between scans. Co-registered MRI and PET-PSMA images would ideally be acquired on a combined PET/MRI scanner, or if not available, within the same week with identical bladder and bowel preparation. Finally, only four subjects in the current cohort underwent radical prostatectomy. Cross-reference to histopathological lesion localisation and Gleason staging was limited to biopsy specimens, which may not reflect true extent or stage of disease. Reference contours by consensus in this setting therefore depend on observer experience and familiarity with the relevant imaging may have affected the results.

Further work could evaluate these findings in a larger prospective cohort and evaluate whether thresholding approaches to delineation could be optimised with the integration of clinical information, such as biopsy Gleason score or PSA level, for the purposes of targeted dose escalation planning. Cross-reference to whole-prostate histopathology specimen for validation of manual and semi-automated contours remains invaluable. Other PSMA radiotracers, such as ^{18}F-PSMA-1007, may better resolve prostate lesions[35,36] and should be investigated with cross-reference to histopathology in the future. Lastly, direct evaluation could be undertaken as to whether the addition of DCE-MRI to bpMRI examinations adds any benefit to cancer delineation when PSMA-PET is available.

In summary, this study found that use of PSMA-PET alone or in conjunction with bpMRI for DIL is associated with lower interobserver variability than bpMRI-based delineations. If available, PSMA-PET should be incorporated into prostate radiotherapy planning for focal radiotherapy escalation, particularly in the setting of smaller lesions which may be relatively occult on DWI. Region of interest segmentation according to ADC thresholding on bpMRI is challenging due to variability in background prostate radiological findings. While quantitative PSMA-PET contouring with $\%SUV_{max}$ appears to be more viable, in both instances, a more sophisticated approach factoring for the effects of lesion volume, location within the prostate, and Gleason grade on lesion-to-background contrast may improve the consistency of semi-automated methods.

REFERENCES

1. Viani GA, Stefano EJ, Afonso SL. Higher-than-conventional radiation doses in localized prostate cancer treatment: a meta-analysis of randomized, controlled trials. *Int J Radiat Oncol Biol Phys* 2009; **74**: 1405–18. doi: https://doi.org/10.1016/j.ijrobp.2008.10.091
2. Pollard JM, Wen Z, Sadagopan R, Wang J, Ibbott GS. The future of image-guided radiotherapy will be Mr guided. *Br J Radiol* 2017; **90**: 20160667. doi: https://doi.org/10.1259/bjr.20160667
3. Lips IM, van der Heide UA, Haustermans K, van Lin ENJT, Pos F, Franken SPG, et al. Single blind randomized phase III trial to investigate the benefit of a focal lesion ablative microboost in prostate cancer (FLAME-trial): study protocol for a randomized controlled trial. *Trials* 2011; **12**: 255. doi: https://doi.org/10.1186/1745-6215-12-255
4. Draulans C, De Roover R, van der Heide UA, Haustermans K, Pos F, Smeenk RJ, et al. Stereotactic body radiation therapy with optional focal lesion ablative microboost in prostate cancer: topical review and multicenter consensus. *Radiother Oncol* 2019; **140**: 131–42. doi: https://doi.org/10.1016/j.radonc.2019.06.023
5. Stoyanova R, Chinea F, Kwon D, Reis IM, Tschudi Y, Parra NA, et al. An Automated Multiparametric MRI Quantitative Imaging Prostate Habitat Risk Scoring System for Defining External Beam Radiation Therapy Boost Volumes. *Int J Radiat Oncol Biol Phys* 2018; **102**: 821–9. doi: https://doi.org/10.1016/j.ijrobp.2018.06.003
6. van Schie MA, Dinh CV, Houdt PJvan, Pos FJ, Heijmink SWTJP, Kerkmeijer LGW, et al. Contouring of prostate tumors on multiparametric MRI: evaluation of clinical delineations in a multicenter radiotherapy trial. *Radiother Oncol* 2018; **128**: 321–6. doi: https://doi.org/10.1016/j.radonc.2018.04.015
7. Borren A, Moman MR, Groenendaal G, Boeken Kruger AE, van Diest PJ, van der Groep P, et al. Why prostate tumour delineation based on apparent diffusion coefficient is challenging: an exploration of the tissue microanatomy. *Acta Oncol* 2013; **52**: 1629–36. doi: https://doi.org/10.3109/0284186X.2013.787164
8. Hofman MS, Lawrentschuk N, Francis RJ, Tang C, Vela I, Thomas P, et al. Prostate-Specific membrane antigen PET-CT in patients with high-risk prostate cancer before curative-intent surgery or radiotherapy (proPSMA): a prospective, randomised, multicentre study. *Lancet* 2020; **395**: 1208–16. doi: https://doi.org/10.1016/S0140-6736(20)30314-7
9. Zamboglou C, Drendel V, Jilg CA, Rischke HC, Beck TI, Schultze-Seemann W, et al. Comparison of ^{68}Ga-HBED-CC PSMA-PET/CT and multiparametric MRI for gross tumour volume detection in patients with primary prostate cancer based on slice by slice comparison with histopathology. *Theranostics* 2017; **7**: 228–37. doi: https://doi.org/10.7150/thno.16638
10. Pepe P, D'Urso D, Garufi A, Priolo G, Pennisi M, Russo G, et al. Multiparametric MRI apparent diffusion coefficient (ADC) accuracy in diagnosing clinically significant prostate cancer. *In Vivo* 2017; **31**: 415–8. doi: https://doi.org/10.21873/invivo.11075
11. Costa DN, Xi Y, Aziz M, Passoni N, Shakir N, Goldberg K, et al. Prospective inclusion of apparent diffusion coefficients in multiparametric prostate MRI structured reports: discrimination of clinically insignificant and significant cancers. *AJR Am J Roentgenol* 2019; **212**: 109–16. doi: https://doi.org/10.2214/AJR.18.19937
12. Lebovici A, Sfrangeu SA, Feier D, Caraiani C, Lucan C, Suciu M, et al. Evaluation of the normal-to-diseased apparent diffusion coefficient ratio as an indicator of prostate cancer aggressiveness. *BMC Med Imaging* 2014; **14**: 15. doi: https://doi.org/10.1186/1471-2342-14-15
13. Nagayama M, Watanabe Y, Terai A, Araki T, Notohara K, Okumura A, et al. Determination of the cutoff level of apparent diffusion coefficient values for detection of prostate cancer. *Jpn J Radiol* 2011; **29**: 488–94. doi: https://doi.org/10.1007/s11604-011-0586-6
14. Esen M, Onur MR, Akpolat N, Orhan I, Kocakoc E. Utility of ADC measurement on diffusion-weighted MRI in differentiation of prostate cancer, normal prostate and

prostatitis. *Quant Imaging Med Surg* 2013; **3**: 210–6. doi: https://doi.org/10.3978/j.issn.2223-4292.2013.08.06
15. Rinaldi D, Fiocchi F, Ligabue G, Bianchi G, Torricelli P. Role of diffusion-weighted magnetic resonance imaging in prostate cancer evaluation. *Radiol Med* 2012; **117**: 1429–40. doi: https://doi.org/10.1007/s11547-012-0832-8
16. Abedi I, Tavakkoli MB, Jabbari K, Amouheidari A, Yadegarfard G. Dosimetric and radiobiological evaluation of multiparametric MRI-guided dose painting in radiotherapy of prostate cancer. *J Med Signals Sens* 2017; **7**: 114–21.
17. 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**: 16–40. doi: https://doi.org/10.1016/j.eururo.2015.08.052
18. Zamboglou C, Schiller F, Fechter T, Wieser G, Jilg CA, Chirindel A, et al. 68)Ga-HBED-CC-PSMA PET/CT Versus Histopathology in Primary Localized Prostate Cancer: A Voxel-Wise Comparison. *Theranostics* 2016; **6**: 1619–28. doi: https://doi.org/10.7150/thno.15344
19. Zamboglou C, Thomann B, Koubar K, Bronsert P, Krauss T, Rischke HC, et al. Focal dose escalation for prostate cancer using ^{68}Ga-HBED-CC PSMA PET/CT and MRI: a planning study based on histology reference. *Radiat Oncol* 2018; **13**: 81. doi: https://doi.org/10.1186/s13014-018-1036-8
20. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol* 2016; **121**: 169–79. doi: https://doi.org/10.1016/j.radonc.2016.09.009
21. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–21. doi: https://doi.org/10.1109/TMI.2004.828354
22. Piert M, Shankar PR, Montgomery J, Kunju LP, Rogers V, Siddiqui J, et al. Accuracy of tumor segmentation from multi-parametric prostate MRI and ^{18}F-choline PET/CT for focal prostate cancer therapy applications. *EJNMMI Res* 2018; **8**: 23. doi: https://doi.org/10.1186/s13550-018-0377-5
23. Rischke HC, Nestle U, Fechter T, Doll C, Volegova-Neher N, Henne K, et al. 3 Tesla multiparametric MRI for GTV-definition of Dominant Intraprostatic Lesions in patients with Prostate Cancer--an interobserver variability study. *Radiat Oncol* 2013; **8**: 183. doi: https://doi.org/10.1186/1748-717X-8-183
24. Steenbergen P, Haustermans K, Lerut E, Oyen R, De Wever L, Van den Bergh L, et al. Prostate tumor delineation using multiparametric magnetic resonance imaging: inter-observer variability and pathology validation. *Radiother Oncol* 2015; **115**: 186–90. doi: https://doi.org/10.1016/j.radonc.2015.04.012
25. Zamboglou C, Fassbender TF, Steffan L, Schiller F, Fechter T, Carles M, et al. Validation of different PSMA-PET/CT-based contouring techniques for intraprostatic tumor definition using histopathology as standard of reference. *Radiother Oncol* 2019; **141**: 208–13. doi: https://doi.org/10.1016/j.radonc.2019.07.002
26. Draulans C, De Roover R, van der Heide UA, Kerkmeijer L, Smeenk RJ, Pos F, et al. Optimal ^{68}Ga-PSMA and ^{18}F-PSMA PET window levelling for gross tumour volume delineation in primary prostate cancer. *Eur J Nucl Med Mol Imaging* 2020;06 Oct 2020. doi: https://doi.org/10.1007/s00259-020-05059-4
27. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Head-To-Head comparison between high- and Standard-b-Value DWI for detecting prostate cancer: a systematic review and meta-analysis. *AJR Am J Roentgenol* 2018; **210**: 91–100. doi: https://doi.org/10.2214/AJR.17.18480
28. Zamboglou C, Wieser G, Hennies S, Rempel I, Kirste S, Soschynski M, et al. Mri versus ^{68}Ga-PSMA PET/CT for gross tumour volume delineation in radiation treatment planning of primary prostate cancer. *Eur J Nucl Med Mol Imaging* 2016; **43**: 889–97. doi: https://doi.org/10.1007/s00259-015-3257-5
29. Giesel FL, Sterzing F, Schlemmer HP, Holland-Letz T, Mier W, Rius M, et al. Intra-Individual comparison of 68Ga-PSMA-11-PET/CT and multi-parametric Mr for imaging of primary prostate cancer. *Eur J Nucl Med Mol Imaging* 2016; **43**: 1400–6. doi: https://doi.org/10.1007/s00259-016-3346-0
30. Perk T, Chen S, Harmon S, Lin C, Bradshaw T, Perlman S, et al. A statistically optimized regional thresholding method (SORT) for bone lesion detection in ^{18}F-NaF PET/CT imaging. *Phys Med Biol* 2018; **63**: 225018: 225018: . doi: https://doi.org/10.1088/1361-6560/aaebba
31. Surov A, Meyer HJ, Wienke A. Correlation between apparent diffusion coefficient (ADC) and cellularity is different in several tumors: a meta-analysis. *Oncotarget* 2017; **8**: 59492–9. doi: https://doi.org/10.18632/oncotarget.17752
32. Boesen L, Chabanova E, Løgager V, Balslev I, Thomsen HS. Apparent diffusion coefficient ratio correlates significantly with prostate cancer Gleason score at final pathology. *J Magn Reson Imaging* 2015; **42**: 446–53. doi: https://doi.org/10.1002/jmri.24801
33. Uprimny C, Kroiss AS, Decristoforo C, Fritz J, von Guggenberg E, Kendler D, et al. ^{68}Ga-PSMA-11 PET/CT in primary staging of prostate cancer: PSA and Gleason score predict the intensity of tracer accumulation in the primary tumour. *Eur J Nucl Med Mol Imaging* 2017; **44**: 941–9. doi: https://doi.org/10.1007/s00259-017-3631-6
34. Tschudi Y, Pollack A, Punnen S, Ford JC, Chang Y-C, Soodana-Prakash N, et al. Automatic detection of prostate tumor habitats using diffusion MRI. *Sci Rep* 2018; **8**: 16801. doi: https://doi.org/10.1038/s41598-018-34916-4
35. Rahbar K, Weckesser M, Ahmadzadehfar H, Schäfers M, Stegger L, Bögemann M. Advantage of ^{18}F-PSMA-1007 over ^{68}Ga-PSMA-11 PET imaging for differentiation of local recurrence vs. urinary tracer excretion. *Eur J Nucl Med Mol Imaging* 2018; **45**: 1076–7. doi: https://doi.org/10.1007/s00259-018-3952-0
36. Giesel FL, Knorr K, Spohn F, Will L, Maurer T, Flechsig P, et al. Detection Efficacy of ^{18}F-PSMA-1007 PET/CT in 251 Patients with Biochemical Recurrence of Prostate Cancer After Radical Prostatectomy. *J Nucl Med* 2019; **60**: 362–8. doi: https://doi.org/10.2967/jnumed.118.212233