Supplementary material for:

PSMASegmentator: Open-source, AI-based software for PSMA PET/CT automatic segmentation and prognostic biomarker extraction

1. Supplementary Methods

1.1. – Imaging datasets

The dataset used for model training combined the 597 PSMA PET/CT scans from the AutoPET-III Challenge [1] with 438 PSMA PET/CTs from a previous, prospective, multicentre clinical trial registered with the Australian New Zealand Clinical Trials Registry (ACTRN12615000608561) [2]. All data were anonymised prior to use in model development and validation. Inter-reader variability of the reference annotations was not assessed because the annotations were obtained from previously curated datasets originating from multiple independent studies and institutions.

As described by Ingrisch et al. [1], the PSMA cohort of the AutoPET-III challenge included pre- and/or post-therapeutic scans for PCa patients, of whom 537 displayed PSMA-avid lesions, while 60 did not, with a mean patient age of 71 (standard deviation 8 years). The images for this dataset were acquired across three scanner types: Siemens (CTI, Knoxville, TN) Biograph 64, Siemens Biograph 20mCT, and GE Discovery 690. The Z-range extended from skull base to mid-thigh, with image acquisitions on average 71 minutes following intravenous injection of ^{18}F-PSMA (mean activity 246 MBq, n=369) or ^{68}Ga-PSMA (mean activity 214 MBq, n=228)[1]. More granular data regarding the tracers administered was not collected, in accordance with the AutoPET Challenge's emphasis on tracer-agnostic performance.

Annotation followed a two-step process of tracer-avid lesion identification via visual assessment and clinical reports, followed by manual segmentation in axial slices. This was by a single reader, whose outputs were reviewed by a radiologist with 5 years' experience in hybrid imaging. Scans derived from LMU, Munich, and UKT, Tübingen [1].

The latter training dataset, described by McCarthy et al. [2], comprised 238 patients (438 scans) with biochemically recurrent (BCR) PCa following prostatectomy (n=152) or radiation therapy (n=86), displaying negative or oligometastatic disease (maximum of 3 lesions) on abdominopelvic CT and bone scintigraphy. All patients underwent a [^{68}Ga]Ga-PSMA-11 PET/CT scan, with n=200 receiving a follow-up scan approximately 6 months later, pending clinical indication.

PET/CT acquisition utilized the Siemens Biograph mCT, with Sir Charles Gairdner Hospital (SCGH, n=381 scans) using a model 64 with the XYZ Gauss 6.00 convolution kernel and PSF+TOF 2i21s iterative reconstruction, while Fiona Stanley Hospital (FSH, n=57 scans) used a model 128 with the 5.00 convolution kernel and OSEM3D+TOF 2i21s reconstruction. Acquisition was 60 minutes post-injection of a standardized, weight-based dose of 2 MBq/kg ^{68}Ga-PSMA-11. Z-range was from mid-thigh to vertex, with the low dose CT using 50mAs on 120kVp. PET components were reconstructed with a voxel resolution of $4.07\times4.07\times2mm^3$, while CT resolutions varied from $0.98\times0.98\times2mm^3$ to $1.52\times1.52\times5mm^3$ [2].

Annotation was via MIM Encore, with higher uptake regions classified according to the E-PSMA guidelines' 5-point scoring system [3]. In addition, a global SUVbw > 3 threshold was applied to the PET scan, and uptake in the seminal vesicles, prostate, and/or prostate bed, indicative of local recurrence, was treated as a single lesion for analysis [2].

Model testing was across internal and external datasets, acquired from different sources and thereby presenting diverse radiotracer, scanner type, date range, and PCa disease extent characteristics.

The first internal test set comprised 105 PET/CTs from 51 mCRPC patients, of whom 97 were imaged with [^{68}Ga]Ga-PSMA-11 and eight with [^{18}F]PSMA-1007 between February 2020 and February 2024. All had received at least one line of prior therapy: RT (N=35), chemotherapy (N=34), ADT (N=32), ARPI (N=30), surgery (N=5). Patients generally had ≥ 5 extracranial, oligoprogressing lesions, with heterogeneous responses to initial therapy. The parameters of image acquisition – scanner type, protocol, reconstruction, resolution – mirror those described previously and by McCarthy et al. for the locally acquired training data [2]. Ethics approval came from the SCGOPHCG HREC (EC00271; reference number ET000679), and access to patient images was approved by the Research Governance Office (RGS5892).

Reference annotations were generated using MIM Encore with WA Health's LesionID workflow, with reference SUV uptake values from ROIs in the right liver lobe and the right atrium (blood pool). A fixed SUV threshold of 3 was applied, with the contours being reviewed against radiology reports and physiological uptake in areas such as the brain, bladder, kidneys, parotids, and injection site being manually deleted.

To this internal testing data, a cohort of 57 mPCa patients from SCGH was added. Fifty-five scans were acquired using ^{68}Ga and three with ^{18}F radionuclides, for a total of 58 scans from November 2019 to October 2025. Whole-body acquisition was via Siemens Biograph scanners, models 64 and 128 mCT and Vision 600 Edge, utilizing XYZ Gauss 5/6.00 convolution kernels with reconstruction via OSEM3D+TOF 2i21s and PSF+TOF 4i5s/2i21s. PET resolutions varied between 200x200, 440x440, and 484x484 mm^2 with slice thicknesses of 3- and 5-mm. Annotation used a similar approach to that described previously for the first test set.

In addition, data from a previous, prospective, single centre trial at SCGH, described by Kendrick et al.[4], was incorporated into the test set. This comprised test-retest scans from 18 mPCa patients (for a total of 36 scans) using [^{68}Ga]Ga-PSMA-11 and [^{18}F]F-PSMA-1007, acquired between July 2020 and October 2021. Image acquisition and reference segmentation followed similar procedures to the previously described local datasets (see [4] for more details).

External testing data used the PSMA PET/CTs from the open-source DEEP-PSMA dataset [5]. This comprised 100 PET/CTs of mCRPC patients, staged using either ^{68}Ga-PSMA-617 or ^{18}F-DCFPyL tracers, and acquired between August 2015 and June 2017 [6,7]. Whole-body acquisition was via GE Discovery 710 & 690, Siemens Biograph and Siemens Vision 600 PET/CTs, with the PET components reconstructed using the EANM EARL-compliant resolution recovery parameters [8]. All PSMA PET/CTs were contoured in MIM using a fixed SUV $\geq$ 3 threshold followed by the manual removal of physiological regions of tracer avidity by an expert NM physician with five or more years' experience [5].

1.2 – Model initialisation

Model training was initialized using the publicly available MultiTalentV2 pretrained checkpoint (available at: https://zenodo.org/records/13753413), which was developed from a diverse collection of medical imaging datasets to provide broad anatomical representation learning. Fine-tuning was then performed using the combined training cohort described in Section 1.1. No randomly initialized model was trained from scratch.

The pretrained checkpoint corresponded to the 'Dataset619_nativemultistem' model released as part of the MultiTalentV2 framework and was used solely for parameter initialization prior to task-specific training.

1.3 – PSMASegmentator pre-processing

Training utilised the standard nnU-Net V2 framework [9], with model parameters initialized according to the framework's default initialization strategy and no transfer learning or pretrained weights employed.

The command-line interface and Python API provide flexible entry points for both direct and Integrated Development Environment- (IDE)-based use. Users can provide input PET/CTs (individual cases or entire datasets) in DICOM or NIfTI format, and configure optional features including RTSTRUCT processing, segmentation expansion, and SUV thresholding. The tool automatically validates input directories, detects image formats, manages model downloading, and coordinates the processing pipeline to ultimately provide the output segmentations and biomarker information. The tool can also be utilized via Docker [10].

Pre-processing converts PSMA PET/CT inputs into prepared NIfTI volumes standardized for nnU-Net inference [9]. For DICOM inputs, the CT and PET series are identified and validated, PET images are converted to SUV (body-weight normalization) using DICOM dose, time, and half-life metadata, and CTs are converted with Hounsfield unit rescaling. Images are reoriented to a standard anatomical frame, and CTs are resampled to the PET grid if size or spacing differ, thereby ensuring voxel-wise alignment for downstream tasks. For NIfTI inputs, CT and PET pairs are detected, copied to a standardized naming scheme, and CT is similarly resampled if needed. When RTSTRUCTs are provided, they are likewise converted to NIfTI using Plastimatch [11] and the Total Tumour Burden (TTB) mask is extracted for ground-truth generation.

1.4 – PSMASegmentator inference

Segmentation inference leverages the nnU-Net V2 framework [9]. The tool initializes a predictor with configurable Test-Time Augmentation (TTA) [12], performs GPU-optimized patch-based inference with Gaussian smoothing for tile stitching, and automatically handles batch processing of multiple cases. The implementation supports both CPU and GPU execution with automatic resource management and provides timing metrics for performance monitoring.

1.5 – Additional information on external survival analysis dataset

At the time of this analysis, this repository contained all patients with public PSMA-PET scan across the state with a diagnosis of prostate cancer between the 1st January 2015 to the 31st December 2024. The analysis cohort, however, only contains patients also from the South Australian Prostate Cancer Clinical Outcomes Collaborative (SA-PCOCC, managed at Flinders Medical Centre) database who have been consented for inclusion in the SA-PCOCC dataset.

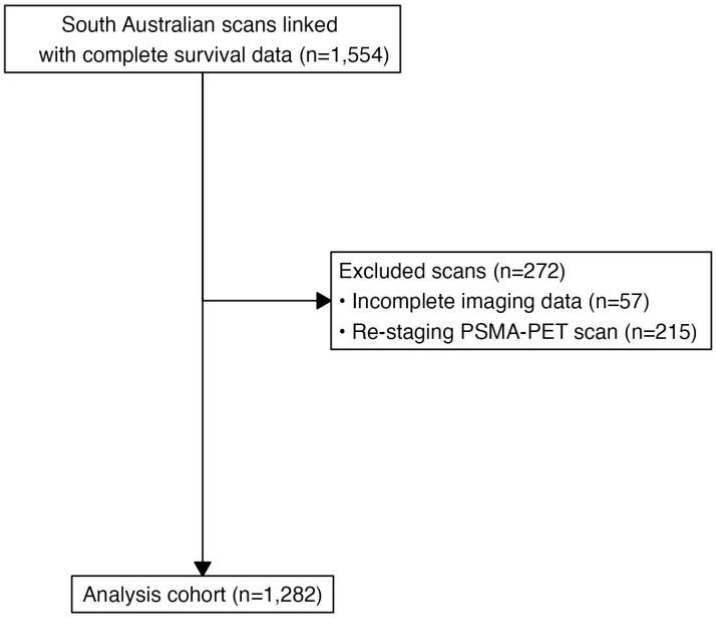

Figure S1 - CONSORT Diagram providing a breakdown of the South Australian dataset used for survival analysis.

1.6 – Missing data in statistical analysis

A complete-case analysis approach was used for survival modelling. Missing data were limited primarily to diagnostic PSA (5.2%) and ISUP Grade Group (13.9%), with no missingness observed for follow-up time or survival outcomes. Thus, there is no anticipated impact of missingness on univariate survival analysis against PSMASegmentator outputs.

2. Supplementary Results

2.1 Cross Validation results

Table S1 – Lesion- and voxel-level performance across each of the five cross validation folds and averaged across all folds. Values are reported as mean, with the standard deviation in

152 *parentheses. For Dice Similarity Coefficient (DSC), median is also reported, with the*

153 *interquartile range in square brackets.*

154

Fold		1	2	3	4	5	All
Lesion-level	Sensitivity	0.6283 (0.4096)	0.6345 (0.4015)	0.7103 (0.3869)	0.6335 (0.3954)	0.6583 (0.4226)	0.6530 (0.4032)
	Precision	0.6246 (0.4047)	0.65 (0.4048)	0.6482 (0.3741)	0.6279 (0.4001)	0.5984 (0.4067)	0.6298 (0.3981)
	F1-score	0.6084 (0.3896)	0.6191 (0.3815)	0.6502 (0.3572)	0.6088 (0.3789)	0.6044 (0.3939)	0.6182 (0.3802)
Voxel-level	Sensitivity	0.5862 (0.3378)	0.5917 (0.3188)	0.6506 (0.3088)	0.6331 (0.3274)	0.6839 (0.3153)	0.6291 (0.3216)
	Precision	0.5393 (0.3556)	0.6426 (0.3391)	0.5569 (0.3432)	0.5503 (0.3507)	0.5103 (0.3483)	0.5599 (0.3474)
	DSC	0.5211 (0.2945) 0.5916 [0.4733]	0.5663 (0.2813) 0.6493 [0.4152]	0.5475 (0.2729) 0.6269 [0.4057]	0.5429 (0.2922) 0.621 [0.4539]	0.5603 (0.2782) 0.6228 [0.3783]	0.5476 (0.2838) 0.6223 [0.4253]

155

2.2 Fast model results

Table S2 – Segmentation performance of the fast model on the testing data. Values are reported as mean, with the standard deviation in parentheses. For DSC, median is also reported, with the interquartile range in square brackets.

Metric	F-18	Ga-68	Internal	External	All
n scans	29	270	199	100	299
Lesion-level					
Sensitivity	0.7698 (0.2364)	0.6487 (0.2128)	0.6834 (0.233)	0.615 (0.1764)	0.6602 (0.2175)
p-value	0.002		0.002		-
Counts*	205/287	9221/15526	5416/8991	4010/6822	9433/15825
Precision	0.765 (0.3027)	0.9077 (0.1821)	0.868 (0.2279)	0.9448 (0.1172)	0.8937 (0.2005)
p-value	0.062		<0.001		-
Counts*	205/240	9221/9613	5416/5757	4010/4096	9433/9861
F1 score	0.7093 (0.2176)	0.7337 (0.1789)	0.7326 (0.1993)	0.7288 (0.1456)	0.7312 (0.1825)
p-value	0.844		0.269		-
Voxel-level					
Sensitivity	0.4994 (0.2403)	0.5515 (0.2127)	0.5456 (0.2289)	0.5482 (0.1882)	0.5455 (0.216)
p-value	0.095		0.973		-
Precision	0.8525 (0.2254)	0.8678 (0.1964)	0.8448 (0.2147)	0.9085 (0.1568)	0.8668 (0.1988)
p-value	0.627		0.003		-
DSC	0.5779 (0.1958)	0.6405 (0.189)	0.6207 (0.1971)	0.6612 (0.1740)	0.6337 (0.1904)
p-value	0.041		0.091		-
Median	0.5916	0.682	0.6514	0.6857	0.6723
IQR	[0.2447]	[0.2531]	[0.2981]	[0.1601]	[0.2592]

SD = Standard Deviation; DSC = Dice Similarity Coefficient; IQR = Inter-Quartile Range.

**For lesion-level sensitivity and precision, global lesion counts (i.e., across the entire cohort)*

are also provided to indicate the number of missed lesions or false positives, respectively.

***For DSC, the median and IQR are provided to account for the skewed distribution.*

2.3 Total lesion count difference

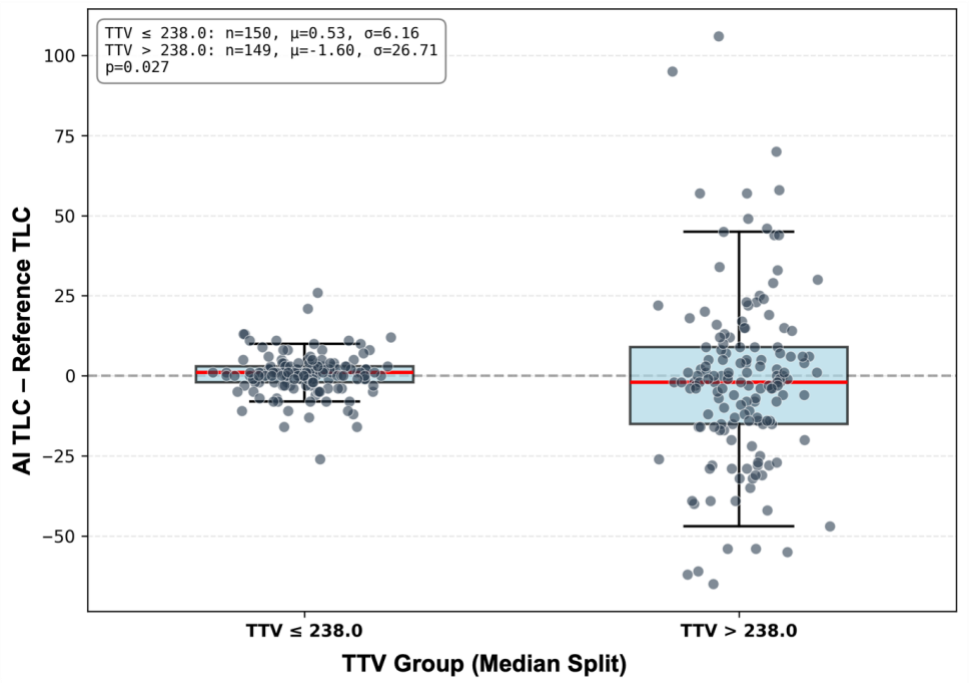

Figure S2 – Boxplots of TLC differences between AI and reference physician segmentations, split above and below the median reference TTV. The text box provides the number of cases in each group, the mean and standard deviation, and the p-value testing statistical difference between the groups. TLC = Total Lesion Count, TTV = Total Tumour Volume.

2.4 Bland-Altman analysis

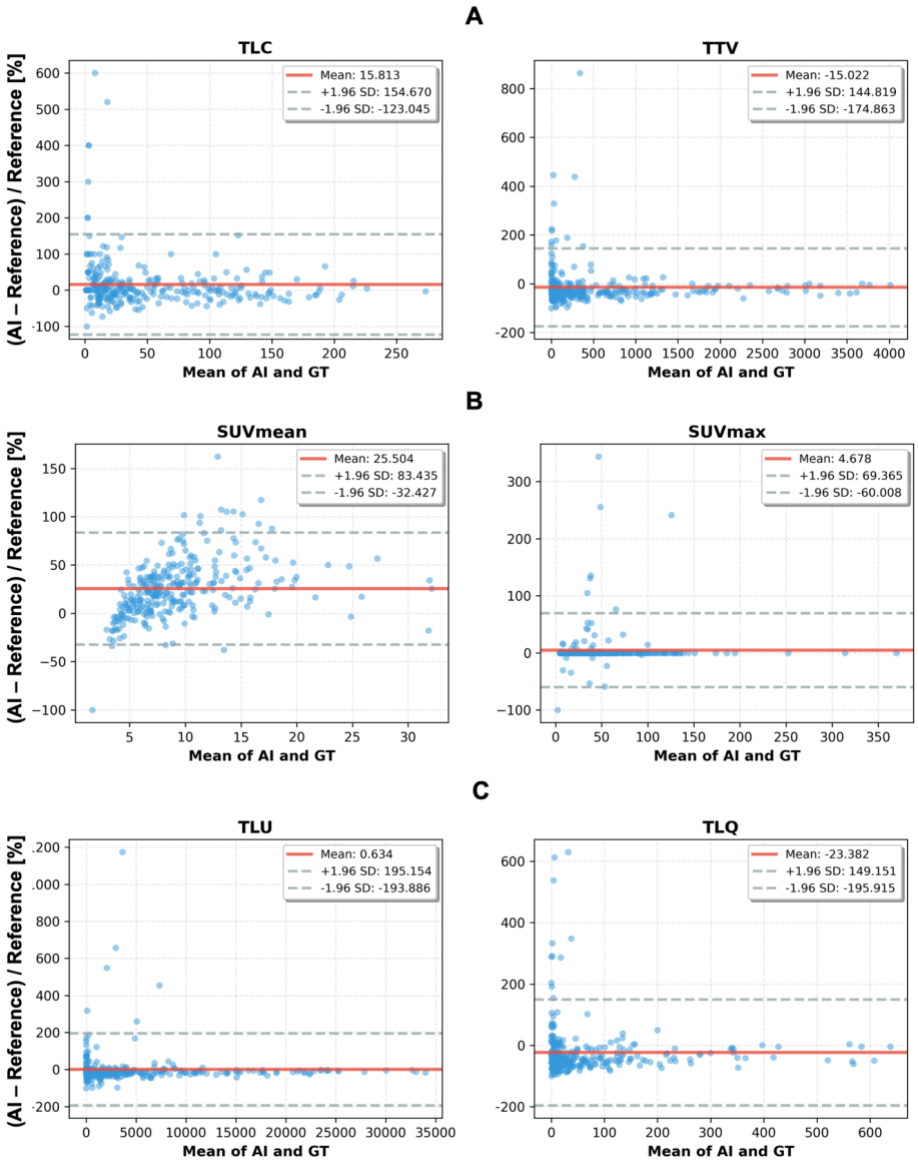

Figure S3 – Bland-Altman plots showing the relative difference in measured biomarkers between measurements for physician and AI-derived segmentations by A) TLC and TTV; B) SUVmean and SUVmax; C) TLU and TLQ. Horizontal lines represent the upper limit of agreement (+1.96 SD), bias (mean difference), and lower limit of agreement (-1.96 SD) from top to bottom. TLC = Total Lesion Count, TTV = Total Tumour Volume, TLU = Total Lesion Uptake, TLQ = Total Lesion Quotient.

4. References

1. Ingrisch, M.*, et al.* Automated lesion segmentation in whole-body pet/ct-multitracer multicenter generalization. in *27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2024)* (2024).
2. McCarthy, M., Francis, R., Tang, C., Watts, J. & Campbell, A. A Multicenter Prospective Clinical Trial of (68)Gallium PSMA HBED-CC PET-CT Restaging in Biochemically Relapsed Prostate Carcinoma: Oligometastatic Rate and Distribution Compared With Standard Imaging. *Int J Radiat Oncol Biol Phys* **104**, 801-808 (2019).
3. Ceci, F.*, et al.* E-PSMA: the EANM standardized reporting guidelines v1.0 for PSMA-PET. *European Journal of Nuclear Medicine and Molecular Imaging* **48**, 1626-1638 (2021).
4. Kendrick, J.*, et al.* Prospective inter- and intra-tracer repeatability analysis of radiomics features in [(68)Ga]Ga-PSMA-11 and [(18)F]F-PSMA-1007 PET scans in metastatic prostate cancer. *Br J Radiol* **96**, 20221178 (2023).
5. Ferdinandus, J.*, et al.* Prognostic biomarkers in men with metastatic castration-resistant prostate cancer receiving [177Lu]-PSMA-617. *Eur J Nucl Med Mol Imaging* **47**, 2322-2327 (2020).
6. Hofman, M.S.*, et al.* [177Lu]-PSMA-617 radionuclide treatment in patients with metastatic castration-resistant prostate cancer (LuPSMA trial): a single-centre, single-arm, phase 2 study. *The Lancet Oncology* **19**, 825-833 (2018).
7. Violet, J.*, et al.* Long-term follow-up and outcomes of retreatment in an expanded 50-patient single-center phase II prospective trial of 177Lu-PSMA-617 theranostics in metastatic castration-resistant prostate cancer. *Journal of Nuclear Medicine* **61**, 857-865 (2020).
8. Aide, N.*, et al.* EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *European Journal of Nuclear Medicine and Molecular Imaging* **44**, 17-31 (2017).
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J. & Maier-Hein, K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203-211 (2021).
10. Combe, T., Martin, A. & Di Pietro, R. To docker or not to docker: A security perspective. *IEEE Cloud Computing* **3**, 54-62 (2016).
11. Sharp, G.C.*, et al.* Plastimatch: an open source software suite for radiotherapy image processing. in *Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR), Amsterdam, Netherlands*, Vol. 3 (2010).
12. Kimura, M. Understanding test-time augmentation. in *International Conference on Neural Information Processing* 558-569 (Springer, 2021).