Physics and Imaging in Radiation Oncology 29 (2024) 100530

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Robustness of magnetic resonance imaging and positron emission tomography radiomic features in prostate cancer: Impact on recurrence prediction after radiation therapy

Arpita Dutta [a], Joseph Chan [b], Annette Haworth [c,*], David J. Dubowitz [d,e], Andrew Kneebone [b], Hayley M. Reynolds [a]

[a] Auckland Bioengineering Institute, The University of Auckland, Auckland, New Zealand
[b] Department of Radiation Oncology, Royal North Shore Hospital, Sydney, New South Wales, Australia
[c] Institute of Medical Physics, School of Physics, University of Sydney, Sydney, New South Wales, Australia
[d] Department of Anatomy and Medical Imaging, Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand
[e] Centre for Advanced MRI, The University of Auckland, Auckland, New Zealand

ARTICLE INFO

Keywords:
Prostate cancer
PET
MRI
Robustness
Radiomics
Recurrence prediction

ABSTRACT

Background and purpose: Radiomic features from MRI and PET are an emerging tool with potential to improve prostate cancer outcomes. However, feature robustness due to image segmentation variations is currently unknown. Therefore, this study aimed to evaluate the robustness of radiomic features with segmentation variations and their impact on predicting biochemical recurrence (BCR).

Materials and methods: Multi-scanner, pre-radiation therapy imaging from 142 patients with localised prostate cancer was used. Imaging included T2-weighted (T2), apparent diffusion coefficient (ADC) MRI, and prostate-specific membrane antigen (PSMA)-PET. The prostate gland and intraprostatic tumours were manually and automatically segmented, and differences were quantified using Dice Coefficient (DC). Radiomic features including shape, first-order, and texture features were extracted for each segmentation from original and filtered images. Intraclass Correlation Coefficient (ICC) and Mean Absolute Percentage Difference (MAPD) were used to assess feature robustness. Random forest (RF) models were developed for each segmentation using robust features to predict BCR.

Results: Prostate gland segmentations were more consistent (mean DC = 0.78) than tumour segmentations (mean DC = 0.46). 112 (3.6 %) radiomic features demonstrated 'excellent' robustness (ICC > 0.9 and MAPD < 1 %), and 480 features (15.4 %) demonstrated 'good' robustness (ICC > 0.75 and MAPD < 5 %). PET imaging provided more features with excellent robustness than T2 and ADC. RF models showed strong predictive power for BCR with a mean area under the receiver-operator-characteristics curve (AUC) of 0.89 (range 0.85–0.93).

Conclusion: When using radiomic features for predictive modelling, segmentation variability should be considered. To develop BCR predictive models, radiomic features from the entire prostate gland are preferable over tumour segmentation-based features.

1. Introduction

Radiomics and artificial intelligence (AI) are emerging tools for implementing precision medicine in prostate cancer [1–3]. Previous studies have demonstrated the potential of pre-treatment MRI or PET-based radiomic features to predict biochemical recurrence (BCR) [4–8], which can affect up to 50 % of patients within ten years of radiation therapy (RT) [3]. Radiomic features also hold promise to assess early post-RT treatment response accurately, complementing the standard blood-based prostate-specific-antigen (PSA) test [1,3]. To be clinically relevant and generalisable, radiomic feature-based predictive models must be robust. Robustness implies stability and reliability, being insensitive to variations in image acquisition, processing, or analysis.

Radiomic features are usually extracted from a segmented region-of-interest (ROI), encompassing the entire prostate gland or the

* Corresponding author at: School of Physics (A28), Physics Road, Camperdown 2006, Australia.
E-mail address: annette.haworth@sydney.edu.au (A. Haworth).

https://doi.org/10.1016/j.phro.2023.100530
Received 28 August 2023; Received in revised form 21 December 2023; Accepted 29 December 2023
Available online 31 December 2023

intraprostatic tumour. In RT, generally, radiation oncologists or radiologists manually segment ROIs using information from multiple imaging modalities (CT, MRI, or PET). However, this is susceptible to inter- and intra-observer variability due to differences in expertise, experience and image contrast. Even with standard PI-RADS guidelines, inter-institutional differences in tumour contouring on MRI cause discrepancies in MRI parameter values from those segmentations [9,10]. In this AI era, the accuracy of automatic segmentation methods depends upon the training data's quality and diversity. Therefore, it is crucial to understand how ROI segmentation variations affect radiomic features and their predictive models.

Previous studies have shown that inter-observer segmentation variations of the prostate or tumour impact feature robustness [11–14]. Most of these studies used single-centre data and didn't explore multi-modality imaging. In real-life, patients often undergo scans at various centres, using different software versions due to upgrades over time. It is known that images acquired from different centres and scanners introduce differences in signal, image reconstruction, and voxel size, leading to feature variations [15,16]. Therefore, this study aimed to investigate the robustness of MRI and prostate-specific membrane antigen (PSMA)-PET radiomic features from manual and automatic segmentations of the prostate and intraprostatic tumours in a heterogeneous pre-RT imaging dataset. Additionally, it examined the impact of segmentation variations on the ability of radiomic features to predict BCR post-RT.

2. Materials and methods

2.1. Data selection

This human research ethics committee (2020/ETH02569) approved study screened a research database from Royal North Shore Hospital (Sydney, Australia) that included prostate cancer patients undergoing definitive RT between January 2010 and December 2019. Inclusion criteria were (1) availability of PSMA-PET/CT and MRI (both T2 and apparent diffusion coefficient (ADC)), (2) curative treatment intent, and (3) biopsy-confirmed adenocarcinoma. Patients with a history of radical prostatectomy (RP) or transurethral resection of the prostate (TURP), or with distant metastatic disease, were excluded. Table 1 summarises the 142 patients' demographic and clinical details, and Fig. 1 shows the study workflow.

2.2. Image acquisition

Multiparametric MRI, including T2-weighted (T2), Diffusion-Weighted Imaging (DWI) and T1-weighted images, were acquired for each patient before RT using various scanners across multiple institutes. MRI scans of 137 patients were acquired using a 1.5 Tesla (T) scanner, and the remaining five patients had MRIs acquired with a 3 T scanner. T2 MRI had an in-plane resolution from 0.3 to 0.6 mm, with a slice thickness from 3.0 to 4.0 mm. DWI images were acquired using at least three b-values (range 50–1400 s/mm^2), in-plane resolution between 0.7 and 1.8 mm and slice thickness from 3.0 to 4.0 mm. Apparent Diffusion Coefficient (ADC) maps were generated on the scanners using the DWI data.

All patients had whole body ^{68}Ga-PSMA-PET/CT scans pre-RT, with 33 scans reconstructed using 3D ordered-subset expectation maximisation, and one using QCFX. The point spread function and time of flight (PSF + TOF) techniques with different iterations (2i21s, 3i21s, 4i10s, 4i21s and 6i10s) were used to reconstruct the remaining scans. A Gaussian filter was applied to all images with a kernel size of 5 mm, except one patient whose kernel size was 4 mm. The PET in-plane resolution varied from 2.2 to 5.5 mm and inter-slice thickness from 2 to 4 mm. PET standardised uptake (SUV) values were computed using the DICOM PET extension in 3D slicer software [17].

Table 1
Patients' demographic and clinical information. BCR – biochemical recurrence. *For BCR prediction, a subset of the cohort with ≥5-year follow-up or BCR were included.

	Robustness Study	BCR Prediction*	
		BCR	Non-BCR
Patients (n)	142	13	36
Age (years)			
Mean ± SD	72.7 ± 7.1	74.5 ± 5.3	71.6 ± 7.1
Range	54–87	66–81	56–87
Initial PSA (ng/mL)	**n (%)**	**n (%)**	**n (%)**
≤ 10	85 (59.9%)	5 (38%)	25 (69%)
>10 & ≤20	38 (26.8%)	4 (31%)	7 (19%)
>20	19 (13.4%)	4 (31%)	4 (11%)
Gleason Grade	**n (%)**	**n (%)**	**n (%)**
3 + 3	1 (0.7%)		
3 + 4	41 (28.9%)		11 (31%)
4 + 3	31 (21.8%)	2 (15%)	11 (31%)
4 + 4	25 (17.6%)	3 (23%)	5 (14%)
4 + 5	38 (26.8%)	7 (54%)	8 (22%)
5 + 4	5 (3.5%)	1 (8%)	1 (3%)
5 + 5	1 (0.7%)		
Pathological T-stage	**n (%)**	**n (%)**	**n (%)**
T1c	31 (21.8%)		12 (33%)
T2	3 (2.1%)		
T2a	12 (8.5%)	3 (23%)	3 (8%)
T2b	21 (14.8%)	1 (8%)	6 (17%)
T2c	15 (10.6%)		3 (8%)
T3	3 (2.1%)	3 (23%)	
T3a	21 (14.8%)	1 (8%)	7 (19%)
T3b	31 (21.8%)	4(31%)	4 (11%)
T4	4 (2.8%)	1 (8%)	1 (3%)
Unknown	1 (0.7%)		
Pathological N-stage	**n (%)**	**n (%)**	**n (%)**
N0	113 (79.6%)	9 (69.2%)	28 (77.8%)
N1	23 (16.2 %)	4 (30.7 %)	5 (13.9 %)
Nx	6 (4.2%)		3 (0.1%)
Hormone treatment	**n (%)**	**n (%)**	**n (%)**
Prior	3 (2.1%)		
Concurrent	72 (50.7%)	8 (61.5%)	22 (61%)
No	67(47.1%)	5(38.5%)	14(38.9%)

2.3. Radiation therapy planning

Patients received various types of RT, including dose-escalated, hypo-fractioned and Stereotactic Body Radiation Therapy (Supplementary Table 1). To reduce the radiation effect on neighbouring organs-at-risk, 61 patients had SpaceOAR Hydrogel inserted between the prostate and rectum before RT.

Planning CT was rigidly co-registered with MRI and whole-body PET/CT using Eclipse treatment planning system (TPS, v16.0; Varian Medical Systems, Paolo Alto, CA, USA). A radiation oncologist with 20 + years' experience manually delineated the prostate gland, excluding the seminal vesicles, as the clinical target volume (CTV) and the tumour as the gross tumour volume (GTV) on planning CT. CTV boundaries were primarily determined using MRI (T1 and T2) due to its superior soft tissue contrast to CT. The GTV was delineated using information from the standard biopsy report, MRI (T2 and ADC) and PET. The GTV and CTV were confirmed by consensus from two other experienced oncologists.

2.4. Automated segmentation

2.4.1. Organ segmentation

The prostate gland was automatically segmented on T2 MRI using a publicly available nnU-Net model, trained using the multi-centre PROMISE12 grand challenge dataset [18,19]. The 2D version of the nnU-Net, which performs slice-by-slice segmentation, was used to generate binary masks of the prostate gland as output.

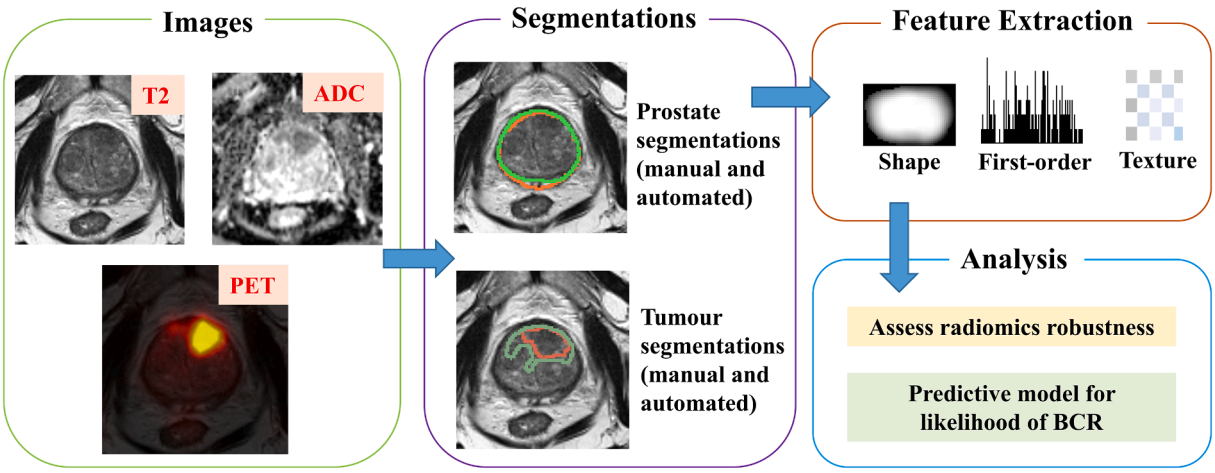

Fig. 1. Workflow used to analyse radiomic feature robustness with different segmentations.

2.4.2. *Tumour segmentation*

Tumours were segmented using two automated methods. First, with an in-house voxel-wise radiomics model, previously developed using data from 19 prostate cancer patients [20–22]. This model used T2, ADC, Dynamic Contrast-Enhanced (DCE) parameters, PSMA-PET images, and clinical features such as age, PSA-level and PET-based features. Since DCE parameters were unavailable in this 142-patient cohort, the model was re-trained on the same 19-patient dataset without them. This re-trained model was used for tumour segmentation.

The second method applied a threshold to the PET standardised uptake (SUV) values. Previous studies have reported various SUV cut-offs with high sensitivity for detecting prostate cancer in ^{68}Ga-PSMA-PET [23,24]. This study conducted a receiver operating characteristic curve and sensitivity–specificity analysis using GTVs as the reference standard to determine the optimal SUV cut-off which was identified as 4.0 (Supplementary Table 2). Automated tumour segmentations were post-processed to remove voxels outside the prostate border and single-voxel predictions.

2.5. *Radiomic feature extraction*

Radiomic features were extracted from each prostate gland and tumour segmentation using 'PyRadiomics' software [25]. Before feature extraction, N4 bias correction was applied to T2 MRI [26], and all images were resampled to modality-specific isotropic voxels (0.3 mm isotropic voxels for T2, 1.25 mm isotropic voxels for ADC maps and 2 mm isotropic voxels for PET). A total of 107 features, including 14 shape-based, 18 first-order statistical features and 75 texture features, were extracted from T2, ADC and PET. Additionally, first-order and texture features were extracted after applying wavelet and Laplacian of Gaussian (LoG) filters, resulting in 930 features from filtered images: 186 from LoG with sigma 3 and 5; 744 from wavelet filtered images. Overall, 1037 features were extracted from each image per patient. Supplementary Table 4 details feature extraction parameters.

2.6. *Statistical analysis*

The Dice coefficient (DC) and mean distance to agreement (MDA) were calculated to assess the agreement between manual (CTV and GTV) and automated segmentations. Radiomic feature robustness was evaluated using the two-way mixed effect model of interclass correlation coefficient (ICC) [27] and mean absolute percent difference (MAPD). ICC calculates the reliability by quantifying the agreement between two feature sets derived from two different segmentations. An ICC value $<$ 0.5 represents poor reliability, between 0.5 and 0.75 indicates moderate reliability, between 0.75 and 0.9 indicates good reliability and $>$ 0.9 represents excellent reliability [28]. MAPD measures the stability of a feature by calculating the average percentage difference between both feature sets, with features having MAPD $<$ 5 % classified as stable [29].

2.7. *Recurrence prediction*

Models were developed to predict BCR after RT using robust radiomic features from pre-RT imaging. BCR was defined as an increase in PSA of 2 ng/mL or more above the nadir PSA post-RT, according to the Phoenix criteria [30]. The non-BCR control group consisted of patients with 5-year BCR-free survival. Out of 142 patients, a subset of 49 patients were included, where 13 (27 %) experienced BCR, and the remaining 36 (73 %) did not (Table 1).

Robust radiomic features with ICC $>$ 0.75 and MAPD $<$ 5 % were used to develop the models. Pearson correlation and minimum Redundancy Maximum Relevance (mRMR) were applied to reduce features, and the top 20 features were selected [31]. The optimal number of features was determined by evaluating ROC-AUC and accuracy metrics for the CTV-based model by differing numbers of features. The feature set was standardised using min–max normalisation. The synthetic minority oversampling technique (SMOTE) method was applied to address the data imbalance caused by the larger number of non-BCR patients [32]. Finally, random forest (RF) models were constructed [33] for each segmentation resulting in five models. The feature selection and model training were performed using 80 % patients' data (n = 39). The model performances were evaluated using sensitivity, specificity, accuracy and area under the receiver-operator-characteristics (ROC) curve (AUC). Due to the small dataset, performance metrics were reported using a stratified 10-fold-cross-validation approach to obtain reliable estimates using the entire dataset.

3. Results

3.1. *Segmentation agreement*

The mean DC between the CTV and the nnU-Net segmentation was 0.78 ± 0.1 (range 0.02 to 0.92), with only seven patients having DC values below 0.6 (Fig. 2D) and mean MDA was 2.45 ± 1.8 (Supplementary Table 3). The mean DC between the GTV and PET thresholding segmentation was 0.46 ± 0.2 (range 0 to 0.96), and mean MDA was 5.6 ± 7.4. The mean DC between the GTV and radiomics model-based tumour segmentations was 0.45 ± 0.2 (range 0 to 0.85), and mean MDA was 4.71 ± 4.7. The radiomics model failed to identify any voxels within the GTV of eight patients, and the PET thresholding failed to identify any voxels within the GTV of twelve patients.

Fig. 2A–C shows the prostate segmentations of three patients. The

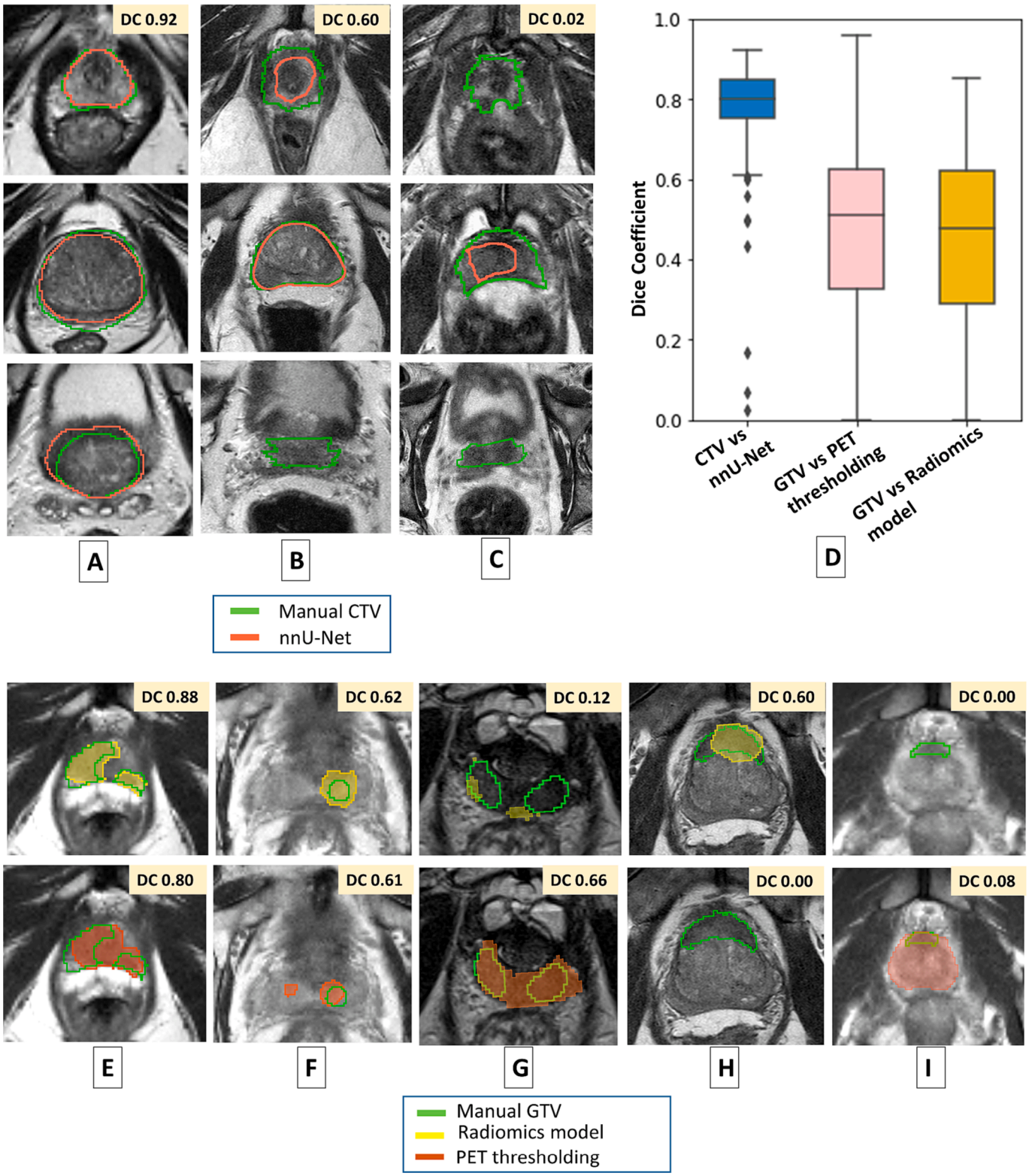

Fig. 2. (A–C) Prostate segmentations of three patients showing the apex (top row), mid gland (middle row), and base (bottom row); (D) boxplots showing the DC for each automated segmentation compared to the manual segmentations; (E–I) tumour segmentations for five patients showing the GTV and the radiomics model-based segmentations (top row), and the GTV and PET thresholding segmentation (bottom row). The corresponding pair-wise DC is provided in the top right corner for (A–C) and (E–I).

patient in Fig. 2A achieved the best segmentation agreement with a DC of 0.92, closely matching at the prostate base, mid-gland, and apex. In Fig. 2B, the segmentation exhibited intermediate agreement (DC = 0.6), while Fig. 2C shows the worst agreement (DC = 0.2) across all patients. Notably, most segmentation mismatches occurred at the apex and base.

Fig. 2E–I display tumour segmentations obtained using the radiomics model (top row) and the PET thresholding method (bottom row) compared to the GTV. The patient in Fig. 2E had the best segmentation

agreement, with both DCs > 0.8. In contrast, the worst-performing segmentation was in the patient in Fig. 2I, as the radiomics model failed to detect the tumour (DC = 0), while PET thresholding over-predicted the tumour (DC = 0.08), identifying the entire prostate due to high tracer uptake (SUV > 4). In Fig. 2G and H, only one of the two automated segmentation methods successfully identified the GTV.

3.2. Radiomic feature robustness

Due to the low agreement between tumour segmentations, the prostate segmentations alone (CTV and nnU-Net) were used to assess radiomic feature robustness. To ensure significant spatial overlap, patients with a DC of 0.6 or higher were considered, giving data from 135 patients. These patients had a mean MDA of 2.16 ± 0.88 which was considered appropriate. Fig. 3 illustrates the distribution of ICC and MAPD values and the mean and 95 % confidence interval for each feature class. Features that exhibited a statistically significant ICC value ($p < 0.05$) were considered for the analysis. The confidence interval brackets tended to be narrower for filtered image features than original image features, indicating increased robustness to variations in segmentations.

Radiomic features were categorised into having excellent, good, moderate or poor robustness based on their ICC and MAPD values (Table 2). A total of 112 (3.6 %) features had excellent robustness, including 51 from PET, 29 from ADC, and 32 from T2 imaging (Supplementary Fig. 1). Approximately 88 % of these features were from wavelet-filtered images, 6 % from log-sigma-filtered images, and 6 % from original images. The excellent features mainly consisted of grey-level texture features, including GLRLM, GLCM, and GLDM. Only the Root Mean Squared features were classified as excellent among the first-order features across all three imaging modalities. In total, 592 (19 %) radiomic features were considered robust as they fell into the excellent or good categories (reported in Supplementary spreadsheet).

Table 2
The number of radiomic features from each imaging modality classified into four categories (excellent, good, moderate and poor) based on ICC and MAPD values.

Robustness	Criteria	Number of radiomic features N (%)		
		PET	ADC	T2
Excellent	ICC > 0.9 and MAPD < 1 %	51 (4.9 %)	29 (2.8 %)	32 (3.1 %)
Good	ICC > 0.75 and MAPD < 5 %	148 (14.3 %)	172 (16.6 %)	160 (15.4 %)
Moderate	ICC > 0.5 and MAPD < 10 %	205 (19.8 %)	214 (20.6 %)	197 (19 %)
Poor	ICC ≤ 0.5 and/or MAPD ≥ 10 %	614 (59.2 %)	601 (57.9 %)	625 (60.3 %)

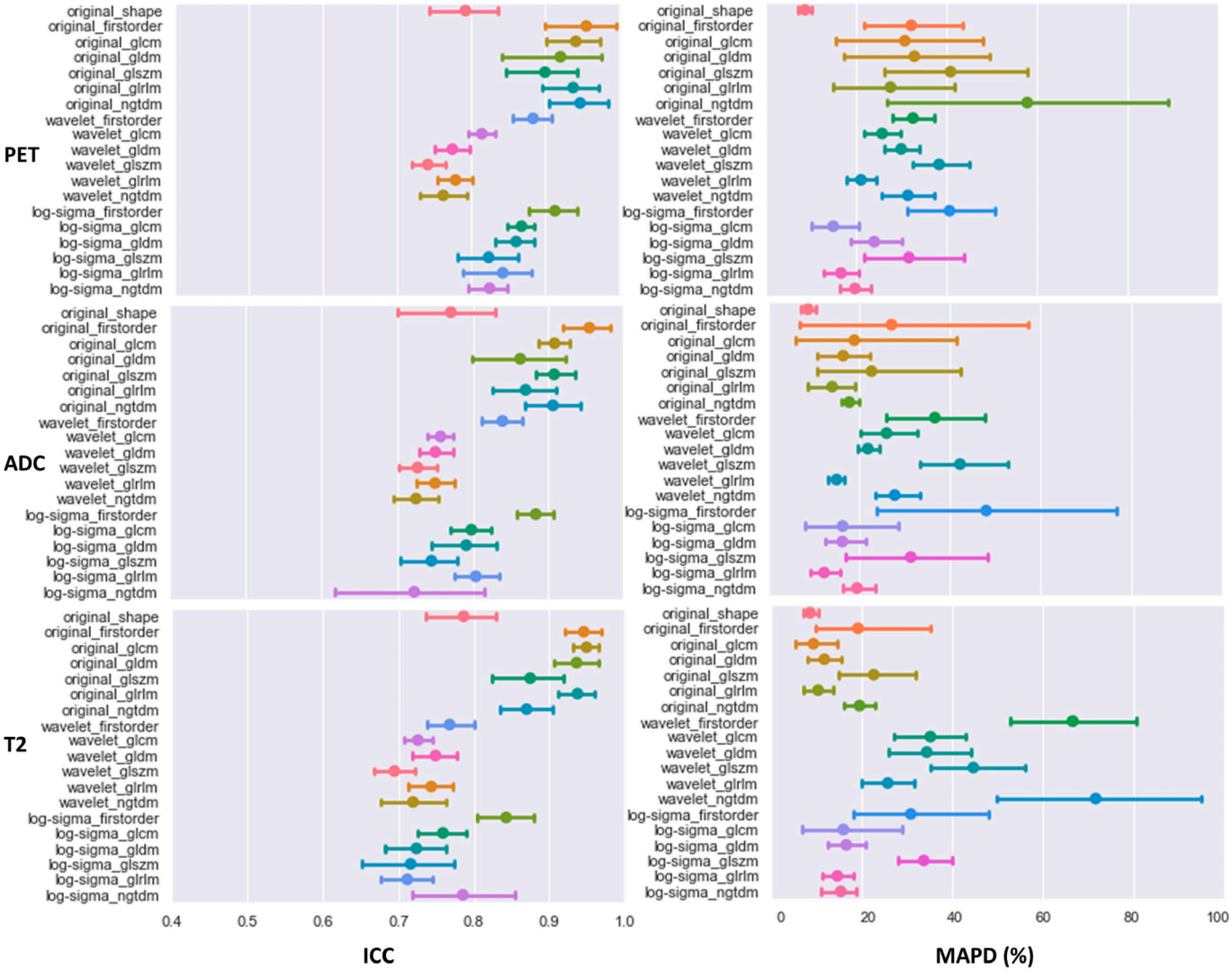

Fig. 3. The ICC and Mean absolute percentage difference (MAPD) along with 95% confidence intervals for every radiomic feature class for T2, ADC and PET.

No shape features were robust (MAPD range 5.5–20.0 %; ICC range 0.23–0.9). Except for the Root Mean Squared features, most first-order features had strong reliability (ICC > 0.9) but low stability (MAPD > 5 %). Kurtosis and skewness were the least robust first-order features with the lowest ICC values (range 0.004–0.88) and MAPD > 10 %. The GLCM feature group contained the most robust features, excluding cluster shade and cluster prominence. Among the GLSZM features, small area emphasis and zone entropy demonstrated robustness, while small area low grey level emphasis did not. Most NGTDM features were not considered robust because their MAPD > 10 %.

3.3. Recurrence prediction

Recurrence prediction models were developed using radiomic features with excellent or good robustness. Patients included had a median follow-up of 4.1 years. The median time for detecting BCR post-RT was 3.2 years (ranging 1.1–5.3 years). The median follow-up period for BCR patients was 5.3 years, while 5.7 years for non-BCR patients.

Table 3 presents the performance metrics for each model. The prostate segmentation models achieved the best performances (AUC 0.91–0.93), outperforming tumour segmentation models (AUC 0.85–0.90). In almost all models, specificity exceeded sensitivity because of the larger sample size of the non-BCR control group.

The most important feature for each model was a texture feature (GLCM or GLDM) derived from PET, and most of the selected twenty features were grey-level textures obtained from filtered images (Supplementary Fig. 2). On average, in each model, 9 were derived from PET scans, 5 from ADC maps, and 6 from T2 MRI.

4. Discussion

This study assessed the robustness of radiomic features with different prostate segmentations and the ability of robust features to predict BCR after RT. To our knowledge, this is the first study evaluating the robustness of prostate radiomic features in both MRI and PET using a real-world multi-scanner dataset, which enhances the generalisability of this study's findings compared to previous studies [11,14,34].

This study's purpose was not to evaluate segmentation performance. Instead, it examined multiple segmentations to understand how their contour variations affect radiomic feature robustness. Prostate segmentations are generally more consistent than tumour segmentations due to prostate cancer's heterogeneous nature. Automated and manual prostate segmentations exhibited good agreement as both were based on T2 MRI, which adequately visualises the prostate boundaries. Automated segmentations performed better in the mid-gland compared to the base or apex regions, consistent with previous studies [19]. One patient with low DC (<0.6) had rectal gas and SpaceOAR, leading to poor segmentation performance, while the other six patients with low DC underwent hormonal therapy, which may have impacted the appearance of the gland on T2 MRI [35]. Poor agreement between some GTVs and PET thresholding or radiomics model segmentations could also be due to patients receiving hormonal therapy (53 % of the cohort), as it influences ADC and PSMA expression [36,37]. Notably, the radiomics-based segmentation model was not trained on post-hormone therapy or multi-scanner images and no publicly available model exists.

Table 3
Performances of recurrence prediction models (average ± SD).

	Segmentation Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
Prostate	CTV	82.0 ± 0.35	94.2 ± 0.12	90.0 ± 0.17	0.93 ± 0.18
	nnU-Net	80.0 ± 0.26	92.0 ± 0.19	86.7 ± 0.17	0.91 ± 0.22
Tumour	GTV	75.0 ± 0.42	97.5 ± 0.08	92.0 ± 0.10	0.88 ± 0.31
	PET thresholding	80.0 ± 0.19	81.8 ± 0.28	80.5 ± 0.19	0.90 ± 0.05
	Radiomics model	70.0 ± 0.48	91.7 ± 0.14	87.0 ± 0.18	0.85 ± 0.32

Ensuring radiomic feature robustness is crucial to building predictive models, encompassing image acquisition, pre-processing, segmentation, and feature extraction. This study follows the guidelines from Image Biomarker Standardization Initiative (IBSI) standards [38,39]. All images were resampled to isotropic voxels before feature extraction to ensure texture features were rotationally invariant. Resampling to modality-specific voxel sizes was done to mitigate up- and down-sampling effects. Additionally, original imaging features were extracted using a modality-specific fixed bin size, while filtered imaging feature extraction employed a fixed bin number. However, Schwier et al. found that bin width doesn't have a strong influence on repeatability of prostate MRI (T2 & ADC)-based radiomic features [15]. Nevertheless, the use of Pyradiomics software [25] deviates slightly from IBSI guidelines due its implementation of image discretization and resampling techniques. Despite identical feature extraction and pre-processing steps used for the source images, the segmentation differences in this study affected the feature robustness. Consistent with our findings, Xue et al. and Urraro et al. reported a small proportion of MRI-based features were reliable [11,13]. Cairone et al. compared T2 MRI features extracted from a 3D prostate volume model using one manual and six semi-automated segmentations, revealing high reliability of grey-level texture features [34]. Similarly, our study found grey-level texture features were more robust than first-order and shape features. While shape features were sensitive to segmentation methods, volume-based and axis length-related shape features had high ICC values, aligned with Cairone et al. [34]. NGTDM features derived from the multi-scanner dataset lacked robustness due to scanner dependency, as did cluster shade and cluster prominence GLCM features.

Currently, the robustness of PET radiomic features has been under-explored. Preliminary findings from the BIOPSTAGE trial indicate that more ADC radiomic features are reliable compared to ^{68}Ga-PSMA-PET radiomic features, though their focus was primarily ISUP grade prediction [14]. Conversely, our study finds that the robustness of PET, ADC, and T2 radiomic features is comparable, with 19.2 %, 19.4 %, and 18.5 % of robust features, respectively. Recent MRI-based studies reported that ADC features are more robust than T2 features against inter-observer segmentations [11,12]. However, we didn't observe a major difference in robustness between T2 and ADC features, potentially due to multi-scanner image use and different b-values for ADC generation. Moreover, bias correction applied to T2 images may have reduced the intensity variations and improved the reliability of T2 features.

Recent evidence suggests that radiomic features extracted from pre-treatment imaging can potentially predict BCR [7,8]. Fernandes et al. developed a BCR prediction model using whole prostate pre-treatment T2 MRI radiomic features and found that this model outperformed standard clinical features or combined models [7]. Gnep et al. found a strong association between T2 texture features and post-RT BCR [8], with all the selected T2 features in these models being grey-level texture features. In our study, the most important feature in all five BCR models was from PET imaging. This may be due to PET SUV's correlation with Gleason grade [24], where higher SUV_{max} values indicate more aggressive and high-grade tumours with a greater risk of recurrence. Papp et al. investigated ^{68}Ga-PSMA-PET-derived radiomic features to predict high versus low-risk prostate tumours and BCR [5]. Similar to our results, they found ^{68}Ga-PSMA-PET features were superior to ADC and T2 for model prediction. Considering the promising performance of the BCR models in this study, prospective trials with long term follow up

data will be needed for clinical validation. Predicting BCR from pre-RT imaging offers an early treatment response indicator, aiding personalized plans and enhanced risk stratification, potentially improving patient outcomes.

This study has several limitations. Firstly, it is constrained by the limitations of automatic tumour segmentations, and a more generalised and robust tumour segmentation approach is necessary to overcome this. Thus, we aim to carry out future research with a larger patient cohort to investigate robustness of radiomic features with varying tumour segmentations. Additionally, manual GTV segmentations based on MRI and biopsy reports are limited because MRI cannot accurately detect tumours smaller than 0.5 cm^3, and these lesions could be missed during biopsy [40]. Secondly, ADC maps were inconsistently generated and with variable b values. Lastly, the cohort size for predicting BCR was small and unbalanced, and didn't incorporate clinical information such as patient demographics, PSA, pathology and comorbities. Therefore, future research will involve a larger patient cohort and consider clinical information in the models.

In conclusion, this study confirmed that variations in prostate gland segmentations on MRI and PET affect the robustness of radiomic features. Due to the high variability between the tumour segmentation methods, radiomic features from the whole prostate offer greater reliability for developing prediction models. Radiomic feature-based models using pre-treatment MRI and PET provide the potential to predict the likelihood of BCR after RT, but further investigation with larger cohorts is needed.

CRediT authorship contribution statement

Arpita Dutta: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Joseph Chan:** Investigation, Data curation, Writing – review & editing. **Annette Haworth:** Conceptualization, Writing – review & editing, Supervision. **David J. Dubowitz:** Writing – review & editing, Supervision. **Andrew Kneebone:** Writing – review & editing. **Hayley M. Reynolds:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Arpita Dutta is supported by the University of Auckland Doctoral Scholarship. Dr Reynolds is funded by a Sir Charles Health Research Fellowship from the Health Research Council of New Zealand. Professor Haworth is supported by a University of Sydney DVCR grant.

Data availability

The datasets used in this study can be shared upon reasonable request from the corresponding author. The supplementary document and spreadsheet contain additional information of this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2023.100530.

References

[1] Stanzione A, Ponsiglione A, Alessandrino F, Brembilla G, Imbriaco M. Beyond diagnosis: is there a role for radiomics in prostate cancer management? Eur Radiol Exp 2023;7:13. https://doi.org/10.1186/s41747-023-00321-4.

[2] O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 2017;14:169–86. https://doi.org/10.1038/nrclinonc.2016.162.

[3] Wang YF, Tadimalla S, Hayden AJ, Holloway L, Haworth A. Artificial intelligence and imaging biomarkers for prostate radiation therapy during and after treatment. J Med Imaging Radiat Oncol 2021;65:612–26. https://doi.org/10.1111/1754-9485.13242.

[4] Zhong QZ, Long LH, Liu A, Li CM, Xiu X, Hou XY, et al. Radiomics of multiparametric MRI to predict biochemical recurrence of localized prostate cancer after radiation therapy. Front Oncol 2020;10:731. https://doi.org/10.3389/fonc.2020.00731.

[5] Papp L, Spielvogel CP, Grubmüller B, Grahovac M, Krajnc D, Ecsedi B, et al. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [68 Ga] Ga-PSMA-11 PET/MRI. Eur J Nucl Med Mol Imaging 2021;48:1795–805. https://doi.org/10.1007/s00259-020-05140-y.

[6] Jia Y, Quan S, Ren J, Wu H, Liu A, Gao Y, et al. MRI radiomics predicts progression-free survival in prostate cancer. Front Oncol 2022;12:974257. https://doi.org/10.3389/fonc.2022.974257.

[7] Fernandes CD, Dinh CV, Walraven I, Heijmink SW, Smolic M, van Griethuysen JJM, et al. Biochemical recurrence prediction after radiotherapy for prostate cancer with T2w magnetic resonance imaging radiomic features. Phys Imaging Radiat Oncol 2018;7:9–15. https://doi.org/10.1016/j.phro.2018.06.005.

[8] Gnep K, Fargeas A, Gutiérrez-Carvajal RE, Commandeur F, Mathieu R, Ospina JD, et al. Haralick textural features on T2-weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. J Magn Reson 2017;45:103–17. https://doi.org/10.1002/jmri.25335.

[9] van Schie MA, Dinh CV, van Houdt PJ, 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;128:321–36. https://doi.org/10.1016/j.radonc.2018.04.015.

[10] Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Update of prostate imaging reporting and data system version 2. Eur Urol 2019;(76):340–51. https://doi.org/10.1016/j.eururo.2019.02.033.

[11] Urraro F, Nardone V, Reginelli A, Varelli C, Angrisani A, Patanè V, et al. MRI radiomics in prostate cancer: a reliability study. Front Oncol 2021;11:805137. https://doi.org/10.3389/fonc.2021.805137.

[12] Wu S, Jiao Y, Zhang Y, Ren X, Li P, Yu Q, et al. Imaging-based individualized response prediction of carbon ion radiotherapy for prostate cancer patients. Cancer Manag Res 2019;11:9121–31. https://doi.org/10.2147/CMAR.S214020.

[13] Xue C, Yuan J, Poon DMC, Zhou Y, Yang B, Yu SK, et al. Reliability of MRI radiomics features in MR-guided radiotherapy for prostate cancer: repeatability, reproducibility, and within-subject agreement. Med Phys 2021;48:6976–86. https://doi.org/10.1002/mp.15232.

[14] Feliciani G, Celli M, Ferroni F, Menghi E, Azzali I, Caroli P, et al. Radiomics analysis on [68 Ga]Ga-PSMA-11 PET and MRI-ADC for the prediction of prostate cancer ISUP grades: preliminary results of the BIOPSTAGE trial. Cancers (Basel) 2022;14:1888. https://doi.org/10.3390/cancers14081888.

[15] Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 2019;9:9441. https://doi.org/10.1038/s41598-019-45766-z.

[16] Guglielmo P, Marturano F, Bettinelli A, Gregianin M, Paiusco M, Evangelista L. Additional value of pet radiomic features for the initial staging of prostate cancer: a systematic review from the literature. Cancers (Basel) 2021;13:6026. https://doi.org/10.3390/cancers13236026.

[17] Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging 2012;30:1323–41. https://doi.org/10.1016/j.mri.2012.05.001.

[18] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.

[19] Litjens G, Toth R, van de Ven W, Hoeks C, Kerkstra S, van Ginneken B, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med Image Anal 2014;18:359–73. https://doi.org/10.1016/j.media.2013.12.002.

[20] Chan TH, Haworth A, Wang A, Osanlouy M, Williams S, Mitchell C, et al. Detecting localised prostate cancer using radiomic features in PSMA PET and multiparametric MRI for biologically targeted radiation therapy. EJNMMI Res 2023;13:34. https://doi.org/10.1186/s13550-023-00984-5.

[21] Reynolds HM, Williams S, Jackson P, Mitchell C, Hofman MS, Hicks RJ, et al. Voxel-wise correlation of positron emission tomography/computed tomography with multiparametric magnetic resonance imaging and histology of the prostate using a sophisticated registration framework. BJU Int 2019;123:1020–30. https://doi.org/10.1111/bju.14648.

[22] Reynolds HM, Williams S, Zhang A, Chakravorty R, Rawlinson D, Ong CS, et al. Development of a registration framework to validate MRI with histology for prostate focal therapy. Med Phys 2015;42:7078–89. https://doi.org/10.1118/1.4935343.

[23] Emmett L, Buteau J, Papa N, Moon D, Thompson J, Roberts MJ, et al. The additive diagnostic value of prostate-specific membrane antigen positron emission tomography computed tomography to multiparametric magnetic resonance imaging triage in the diagnosis of prostate cancer (PRIMARY): a prospective multicentre study. Eur Urol 2021;80:682–9. https://doi.org/10.1016/j.eururo.2021.08.002.

[24] Jiao J, Kang F, Zhang J, Quan Z, Wen W, Zhao X, et al. Establishment and prospective validation of an SUVmax cutoff value to discriminate clinically

significant prostate cancer from benign prostate diseases in patients with suspected prostate cancer by 68Ga-PSMA PET/CT: a real-world study. Theranostics 2021;11: 8396–411. https://doi.org/10.7150/THNO.58140.
[25] Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
[26] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20. https://doi.org/10.1109/TMI.2010.2046908.
[27] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420. https://doi.org/10.1037/0033-2909.86.2.420.
[28] Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
[29] Schmidt RM, Delgadillo R, Ford JC, Padgett KR, Studenski M, Abramowitz MC, et al. Assessment of CT to CBCT contour mapping for radiomic feature analysis in prostate cancer. Sci Rep 2021;11:22737. https://doi.org/10.1038/s41598-021-02154-w.
[30] Roach M, Hanks G, Thames H, Schellhammer P, Shipley WU, Sokol GH, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO phoenix consensus conference. Int J Radiat Oncol Biol Phys 2006; 65:965–74. https://doi.org/10.1016/j.ijrobp.2006.04.029.
[31] Zhao Z, Anand R, Wang M. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. IEEE 6th International Conference on Data Science and Advanced Analytics (DSAA); 2019 Oct; IEEE 2019. p. 442–52. https://doi.org/10.1109/DSAA.2019.00059.
[32] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[33] Ho TK. Random Decision Forests. Proceedings of the 3rd international conference on document analysis and recognition; 1995 Aug 14; Montreal, QC, Canada. IEEE 1995. p. 278-82. https://doi.org/10.1109/ICDAR.1995.598994.
[34] Cairone L, Benfante V, Bignardi S, Marinozzi F, Yezzi A, Tuttolomondo A, et al. Robustness of radiomics features to varying segmentation algorithms in magnetic resonance images. In: International Conference on Image Analysis and Processing. Springer; 2022. p. 462–72. https://doi.org/10.1007/978-3-031-13321-3_41.
[35] Chen M, Hricak H, Kalbhen C, Kurhanewicz J, Vigneron D, Weiss J, et al. Hormonal ablation of prostatic cancer: effects on prostate morphology, tumor detection, and staging by endorectal coil MR imaging. AJR 1996;166:1157–63. https://doi.org/10.2214/ajr.166.5.8615261.
[36] Onal C, Guler OC, Torun N, Reyhan M, Yapar AF. The effect of androgen deprivation therapy on 68Ga-PSMA tracer uptake in non-metastatic prostate cancer patients. Eur J Nucl Med Mol Imaging 2020;47:632–41. https://doi.org/10.1007/s00259-019-04581-4.
[37] Björeland U, Nyholm T, Jonsson J, Skorpil M, Blomqvist L, Strandberg S, et al. Impact of neoadjuvant androgen deprivation therapy on magnetic resonance imaging features in prostate cancer before radiotherapy. Phys Imaging Radiat Oncol 2021;17:117–23. https://doi.org/10.1016/j.phro.2021.01.004.
[38] Depeursinge A, Andrearczyk V, Whybra P, van Griethuysen J, Müller H, Schaer R, et al. Standardised convolutional filtering for radiomics. arXiv:2006.05470v8. 2020. Available from: http://arxiv.org/abs/2006.05470.
[39] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295: 328–38. https://doi.org/10.1148/radiol.2020191145.
[40] Chen MY, Woodruff MA, Dasgupta P, Rukin NJ. Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. Cancer Med 2020;9:7172–82. https://doi.org/10.1002/cam4.3386.