European Journal of Nuclear Medicine and Molecular Imaging (2022) 49:527–538
https://doi.org/10.1007/s00259-021-05430-z

ORIGINAL ARTICLE

The added value of PSMA PET/MR radiomics for prostate cancer staging

Esteban Lucas Solari[1] · Andrei Gafita[1] · Sylvia Schachoff[1] · Borjana Bogdanović[1] · Alberto Villagrán Asiares[1] · Thomas Amiel[2] · Wang Hui[1] · Isabel Rauscher[1] · Dimitris Visvikis[3] · Tobias Maurer[4] · Kristina Schwamborn[5] · Mona Mustafa[1] · Wolfgang Weber[1] · Nassir Navab[6] · Matthias Eiber[1] · Mathieu Hatt[3] · Stephan G. Nekolla[1]

Received: 14 January 2021 / Accepted: 24 May 2021 / Published online: 13 July 2021

Abstract

Purpose To evaluate the performance of combined PET and multiparametric MRI (mpMRI) radiomics for the group-wise prediction of postsurgical Gleason scores (psGSs) in primary prostate cancer (PCa) patients.

Methods Patients with PCa, who underwent [^{68} Ga]Ga-PSMA-11 PET/MRI followed by radical prostatectomy, were included in this retrospective analysis (n = 101). Patients were grouped by psGS in three categories: ISUP grades 1–3, ISUP grade 4, and ISUP grade 5. mpMRI images included T1-weighted, T2-weighted, and apparent diffusion coefficient (ADC) map. Whole-prostate segmentations were performed on each modality, and image biomarker standardization initiative (IBSI)-compliant radiomic features were extracted. Nine support vector machine (SVM) models were trained: four single-modality radiomic models (PET, T1w, T2w, ADC); three PET + MRI double-modality models (PET + T1w, PET + T2w, PET + ADC), and two baseline models (one with patient data, one image-based) for comparison. A sixfold stratified cross-validation was performed, and balanced accuracies (bAcc) of the predictions of the best-performing models were reported and compared through Student's t-tests. The predictions of the best-performing model were compared against biopsy GS (bGS).

Results All radiomic models outperformed the baseline models. The best-performing (mean ± stdv [%]) single-modality model was the ADC model (76 ± 6%), although not significantly better ($p > 0.05$) than other single-modality models (T1w: 72 ± 3%, T2w: 73 ± 2%; PET: 75 ± 5%). The overall best-performing model combined PET + ADC radiomics (82 ± 5%). It significantly outperformed most other double-modality (PET + T1w: 74 ± 5%, $p = 0.026$; PET + T2w: 71 ± 4%, $p = 0.003$) and single-modality models (PET: $p = 0.042$; T1w: $p = 0.002$; T2w: $p = 0.003$), except the ADC-only model ($p = 0.138$). In this initial cohort, the PET + ADC model outperformed bGS overall (82.5% vs 72.4%) in the prediction of psGS.

Conclusion All single- and double-modality models outperformed the baseline models, showing their potential in the prediction of GS, even with an unbalanced cohort. The best-performing model included PET + ADC radiomics, suggesting a complementary value of PSMA-PET and ADC radiomics.

Keywords Prostate cancer · PET/MRI · PSMA · Radiomics · Gleason score

Abbreviations

ADC	Apparent diffusion coefficient
bAcc	Balanced accuracy
bGS	Biopsy Gleason score
DWI	Diffusion weighted image
FLAB	Fuzzy-logically adaptive Bayesian
GGG	ISUP Gleason Grade Groups
GS	Gleason score
IBSI	Image biomarker standardization initiative
IQR	Interquartile range
ISUP	International Society of Urological Pathology
mpMRI	Multiparametric MRI
MRI	Magnetic resonance imaging
PCa	Prostate cancer
PET	Positron emission tomography
PSA	Prostate-specific antigen

Esteban Lucas Solari and Andrei Gafita contributed equally

Mathieu Hatt and Stephan G. Nekolla joint senior authorship

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence).

✉ Esteban Lucas Solari
elucas.solari@tum.de

Extended author information available on the last page of the article

psGS	Postsurgical Gleason score
PSMA	Prostate-specific membrane antigen
RBF	Radial basis function
RFE	Recursive feature elimination
RP	Radical prostatectomy
stdv	Standard deviation
SMOTE	Synthetic minority oversampling technique
SUV	Standard uptake value
SVM	Support vector machine
T1w	T1-weighted MR image
T2w	T2-weighted MR image
VOI	Volume of interest

Introduction

Prostate cancer (PCa) is a leading cause of cancer-associated morbidity and mortality in men [1]. Diagnosis of PCa is commonly achieved by ultrasound-guided needle biopsy and can be improved by multiparametric magnetic resonance imaging (mpMRI) [2, 3]. Positron emission tomography (PET) imaging with PCa-specific tracers can help the delineation of suspicious lesions for guiding repeated biopsies or to improve the sensitivity of lesion detection [2, 4]. More recently, ^{68} Ga-radiolabelled prostate-specific membrane antigen PET (PSMA-PET) demonstrated superiority over other imaging modalities and PET radiotracers in localizing primary staging and biochemical recurrent PCa [5–7]. Moreover, [^{68} Ga]Ga-PSMA-11 PET/MRI showed promising results in aiding targeted biopsy after a previous negative biopsy in patients with high suspicion of PCa [8–10].

Patients with histologically confirmed PCa are initially stratified into risk groups according to serum prostate-specific antigen (PSA) levels, histological findings, and digital-rectal examination results [4]. The Gleason score (GS) extracted from biopsy results or after radical prostatectomy (RP) is the main tool for prognosis, and an indicator of the aggressiveness of PCa. Recently, the International Society of Urological Pathology (ISUP) reached a consensus regrouping of the GS into 5 Gleason Grade Groups (GGG) [11], according to their correlation with patient outcome.

However, in approximately one-third of the patients, biopsy GS (bGS) is different from the final GS determined after surgery (postsurgical GS, psGS), with biopsies tending to underestimate cancer aggressiveness [12]. These discrepancies between the two GS can have important implications in patient management. Therefore, accurate determination of PCa aggressiveness by adding pre-therapeutic imaging features is of high clinical interest.

To achieve this goal, data-driven strategies received much interest in the last decade. In this work, we focus on radiomics, which is the extraction of image features from medical images and their use to build models for improved decision support. Hand-crafted radiomic features have been previously applied for aiding detection and prognosis in breast cancer [13], lung cancer [14], and glioma [15]. MRI-only radiomics have also been applied in PCa prognosis, using GS as a proxy [16–18]. PSMA-PET radiomics [19] and other PET tracers [20] have been independently applied in the discrimination between low- to intermediate-risk (GS $\leq$ 7 or GGG 1–3) and high-risk (GS $\geq$ 8 or GGG 4–5) PCa.

In this study, we investigated the performance of hand-crafted radiomic features extracted from pre-therapeutic [^{68} Ga]Ga-PSMA-11 PET/MRI in predicting psGS in three categories (GGG 1–3, GGG 4, GGG 5). The selection of the three categories was based on a compromise between a more comprehensive prediction, knowing that all the selected Gleason categories represent different clinical outcomes [21], and the availability of data, since there were not enough patient data to represent every Gleason category.

The complementary value of PET and MRI radiomics was evaluated by comparing single- (PET or MRI) and double-modality (PET + MRI) radiomics. In addition, the performance of image-based models was compared to two baseline models: the first one trained with clinical patient-data only, and the second one trained with volume and maximum intensity radiomic features only. Finally, psGS predictions from the best-performing model were compared to assuming bGS.

Methods

Patient population

Patients with histopathologically proven primary adenocarcinoma of the prostate who (i) received a [^{68} Ga]Ga-PSMA-11 PET/MRI at our institution between November 7, 2012, and February 13, 2014, for initial staging of PCa, (ii) had undergone RP, and (iii) had available surgery-obtained GS were included in this retrospective analysis. Out of the 132 screened patients, 101 met the eligibility criteria and were included in this study, whereas 31 patients did not include all the necessary MR images and were excluded. Patient characteristics are summarized in Table 1. Clinical and histopathological information were extracted from hospital database. All patients provided written informed consent for data evaluation and publication. The retrospective data analysis was approved by the medical ethics committee of the Technical University of Munich (reference number: 5665/13S).

The histopathology data were extracted from RP pathology reports. GS values were patient-based, meaning that the total GS per patient was selected, consisting of the sum of the scores of the two most dominant Gleason patterns.

Patients were grouped by GS into three categories: lower than 8 (GGG 1–3), equal to 8 (GGG 4), and higher than 8 (GGG 5).

Imaging protocol

Imaging was performed on an integrated whole-body PET/MRI system (Biograph mMR, Siemens Healthineers, Erlangen, Germany) with a 3 T MRI system. PET images were obtained after intravenous injection of a median of 142.0 (interquartile range [IQR]: 118.3–156.8) MBq of [^{68} Ga] Ga-PSMA-11 synthesized as previously described [22]. Twenty milligrams of furosemide were injected right after tracer administration. PET/MRI acquisitions started at a median of 60.4 (IQR: 51.5–73.2) minutes following radiopharmaceutical injection. Subsequently, mpMRI examination of the prostate was performed simultaneously within a 15-min single bed position PET scan (PET), including a coronal T1-weighted image (T1w), an isotropic T2-weighted image (T2w), and an axial apparent diffusion coefficient map (ADC), all centered on the prostate. All acquisition and reconstruction protocols for both PET and MRI sequences were the same for all patients included in the present study (Supplementary Table 1).

Image segmentation

For the extraction of radiomic features, volumes of interest (VOI) in PSMA-PET and MRI images were first individually segmented. To avoid the limitations of radiomics for small lesions in PET images [23] and the complexities of multi-lesion characterization through hand-crafted radiomic features, whole-prostate segmentations were performed. PSMA-PET images were segmented using a previously validated fuzzy-logically adaptive Bayesian (FLAB) segmentation tool [24], which provides accurate estimation of volumes of interest through modelling of noise and blur characteristics of PET imaging [25]. Whole prostates from MR images (T1w, T2w, and ADC maps) were manually segmented in each modality. The segmentations were performed by a nuclear medicine physician with 3 years of experience in PSMA hybrid imaging.

Table 1 Patient cohort characteristics. Data are median (interquartile range) or n (%); *PSA*, prostate-specific antigen; *RP*, radical prostatectomy; *Missing biopsy results for 30 patients (n = 71)

	All eligible patients (n=101)
Age (years)	68 (63-73)
Weight (kg)	84 (77-95)
Gleason score (biopsy) *	
<8 (GGG 1-3)	31 (44%)
=8 (GGG 4)	19 (27%)
>8 (GGG 5)	21 (30%)
Gleason score (RP)	
<8 (GGG 1-3)	60 (59%)
=8 (GGG 4)	23 (23%)
>8 (GGG 5)	18 (18%)
initial PSA (ng/ml)	12 (7.3-28.1)

Radiomic features extraction

The radiomic features were extracted from the segmented volumes, in accordance to the image biomarker standardization initiative (IBSI) guidelines [26]. Two different discretization approaches were used. To preserve the original intensity scale and meaning of the voxel values, quantitative functional imaging modalities (PSMA-PET and ADC maps) were discretized using fixed bin width (FBW) sizes (bin sizes PET [SUV] = 0.030, 0.060, 0.125, 0.250, 0.500, 1.000; bin sizes ADC [10^{-6} mm^2/s] = 10, 25, 50, 100, 200, 400). According to the IBSI guidelines, in order to normalize the images and prioritize contrast inside the VOIs, the discretization of non-quantitative MRI T1w and T2w images was performed using fixed bin numbers (FBNs, number of bins = 8, 16, 32, 64, 128, 256) discretization. From these discretization schemes, the best-performing one for each model based on its balanced accuracy was selected as the final model.

Overall, 107 3D radiomic features were extracted from the original VOIs without resampling, using PyRadiomics [27], which included: first order (n = 18), shape (n = 14), Gray Level Co-occurrence Matrix (GLCM) (n = 24), Gray Level Size Zone Matrix (GLSZM) (n = 16), Gray Level Run Length Matrix (GLRLM) (n = 16), Neighbouring Gray Tone Difference Matrix (NGTDM) (n = 5), and Gray Level Dependence Matrix (GLDM) (n = 14). The feature extraction workflow is described in Fig. 1.

Based on previous works [28–31], six features derived from commonly used PSMA-PET quantitative biomarkers were also included among the PSMA-PET features: SUV_{peak} (maximum average SUV within a 1-cm^3 spherical volume), relative SUV_{peak} (ratio of SUV_{peak} and the mean SUV of the VOI), volume of the 40% of SUV_{max} isocontour [40% Volume], volume fraction of the 40% isocontour [40% Fraction], SUV_{mean} in the 40% isocontour [40% SUV_{mean}], and "Total SUV" (product of the 40% SUV_{mean} and the 40% Volume). The isocontour volume and the "Total SUV" features were based on "PSMA-ligand tumor volume" [PSMA-TV] and "PSMA-ligand total lesion" [PSMA-TL] [28], adapted to our prostate segmentation.

All the features used in this work are described in the Supplementary Table 2.

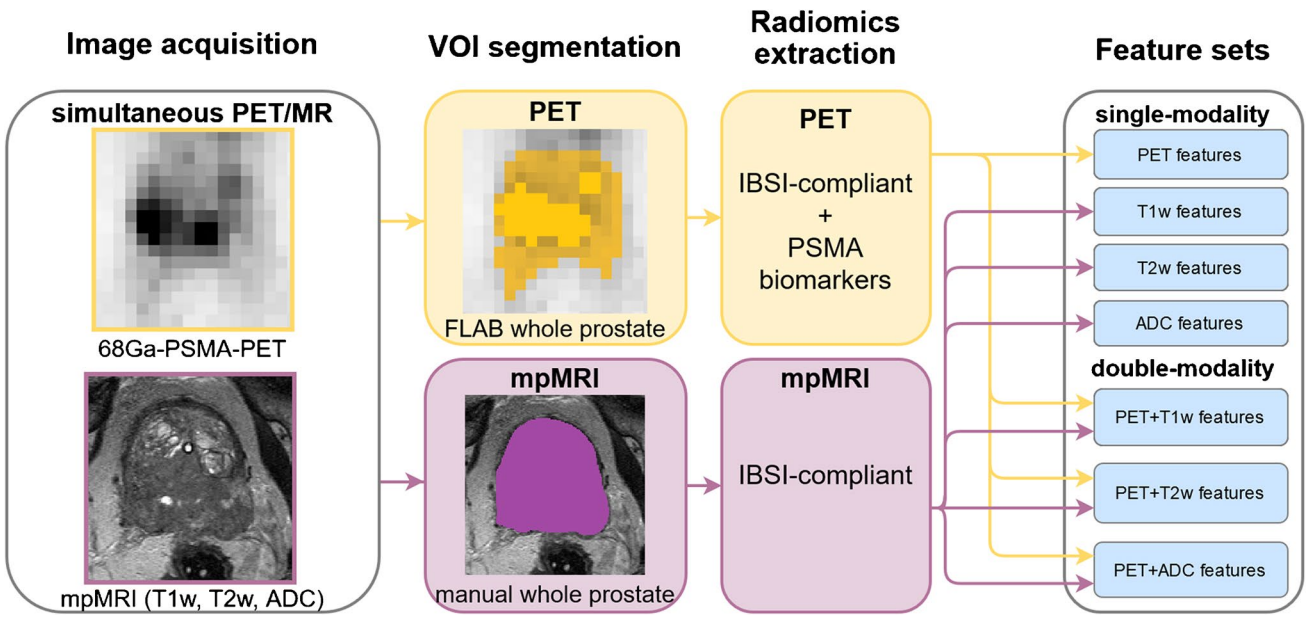


Fig. 1 Feature extraction workflow for PSMA-PET and mpMRI images

Prediction models

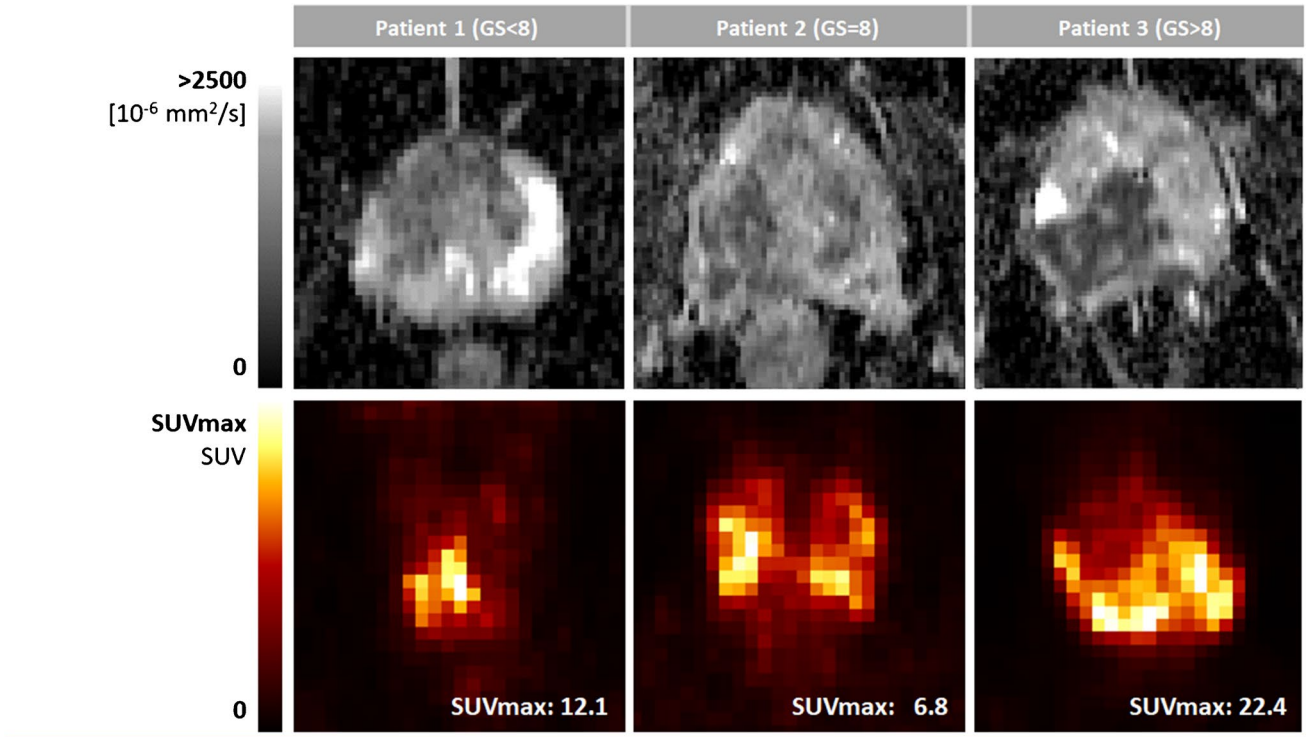
A machine learning model was trained to identify GS using a 3-class support vector machine (SVM) with a radial basis function (RBF) kernel and a "one-vs-rest" multi-class approach. The SVM was trained with up to 10 previously selected PET and MRI radiomic features, using a recursive feature elimination (RFE) method for the selection of the radiomic signature. The cohort was split into training and validation sets using a sixfold stratified cross-validation, with a 2:1 patient ratio for training (n = 67) and validation (n = 34) in each fold, respectively. As examples of radiomic signatures, the three most relevant features of two models (with ADC and PET radiomics, respectively) are presented in Fig. 2 for three example patients.

Given the strong class imbalance (59% of the patients belong to the class GGG 1–3), a method to balance the training datasets is required. There are several alternatives, which usually imply an oversampling of the less prevalent classes, an undersampling of the most prevalent class, or a combination of both. With a limited dataset, our study is better suited for oversampling techniques, to avoid neglecting important information. The synthetic minority oversampling technique (SMOTE) was applied to oversample all training features in both less prevalent classes (GGG 4 and GGG 5) up to a 1:1 proportion with the most prevalent class (GGG 1–3). The augmented training data was used to train the SVM for each of the 6 cross-validation cycles. The trained models were then tested in the prediction of GS with the non-augmented validation data. The implemented training and validation of the models is displayed in Fig. 3.

Separate models were trained using either radiomic features from a single image type (single-modality models: PSMA-PET, T1w, T2w, ADC) or combined from PET and each MR sequence (double-modality models: PSMA-PET + T1w, PSMA-PET + T2w, PSMA-PET + ADC). The use of more than two image types — such as PET and several MR sequences — was not implemented for two reasons: first, the exponentially higher computing time required to train a model with all the features and combined hyperparameters from many modalities; second, the relative low outcome changes obtained from adding extra modalities in previous experiences [32].

We compared our radiomic models against two ad hoc baseline SVM models to evaluate their performance. With these comparisons, we investigated if our radiomics performed better than using other available information. To study the added value of image radiomics beyond conventional clinical information, a patient-data baseline was established. To confirm that our models did not only rely on surrogates of volume or maximum intensity voxel values, an image-based model ("radiomics baseline") was trained. This echoes previous studies [23, 33] which suggest that, if left unchecked, some radiomics signatures may be little more than proxies for simpler statistics, like the number of voxels of the VOI (i.e., its volume). Adding such a baseline model helps us discard this hypothesis whenever our models outperform the radiomics baseline, implying that our models rely on more complex features.

For the training of these two baseline models, we followed the same workflow as in the image radiomics models (Fig. 3), but instead of starting with a set of image radiomic



Radiomic signature ADC (first 3 features only)			
1. GLCM: Autocorrelation	139.7	107.9	82.0
2. GLSZM: Small Area High Gray Level Emphasis	62.3	63.8	70.7
3. GLDM: Small Dependence High Gray Level Emphasis	20.7	15.8	11.7
Radiomic signature PET (first 3 features only)			
1. GLDM: Dependence Entropy	7.74	7.88	8.89
2. NGTDM: Contrast	0.59	0.52	1.93
3. GLCM: Contrast	507.5	226.4	3857.6

Fig. 2 Examples of radiomic signatures (three most relevant features only) of three patients, one per GS category

features, we used: only relevant patient data (i.e., age, weight, and initial PSA [iPSA]) for the patient-data baseline; only VOI volume and maximum values (i.e., PET VOI volume, SUV_{max}, and ADC_{max}) for the radiomics baseline.

Statistical analysis

Since the dataset was unbalanced and classification performance can be different for each of the 3 classes, the results are expressed as the balanced accuracy (bAcc) among the three classes (Eq. 1):

$$bAcc = \frac{1}{3}\left(\frac{TP_{GGG1-3}}{(TP+FN)_{GGG1-3}} + \frac{TP_{GGG4}}{(TP+FN)_{GGG4}} + \frac{TP_{GGG5}}{(TP+FN)_{GGG5}}\right); \quad (1)$$

where TP_{CLASS} is the number of True Positive classifications for the class CLASS, and FN_{CLASS} is the number of False Negative classifications for the class CLASS.

The results from the GS predictions are presented as the sixfold mean and standard deviations of bAcc, sensitivities, and specificities of the best prediction using the validation data. Results from the baseline models, PET-only radiomics, each MR-modality (T1w, T2w, ADC) radiomics only, and combined PET + MR radiomics were compared using Student's t-test (statistical significance: $p < 0.05$; normality

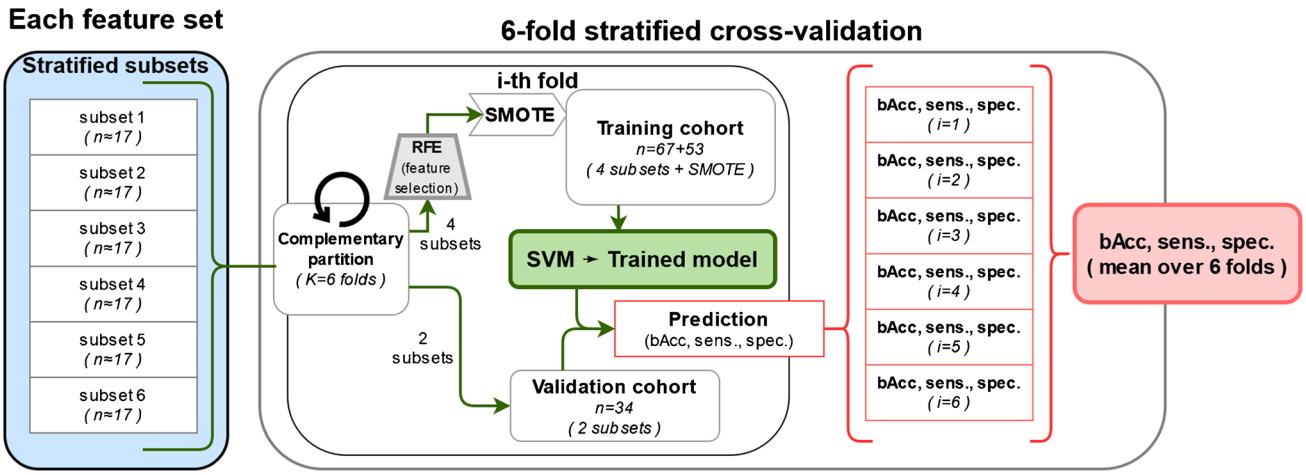


Fig. 3 Training and validation workflow of each SVM model for the prediction of GS

test: Shapiro–Wilk), to evaluate the complementary value of radiomics from both imaging modalities.

To assess the potential clinical value of using PET-MR radiomics in the prediction of GS, we compared the use of biopsy GS (bGS) as a substitute for postsurgical GS against the predictions from our best-performing model. All the predictions from our best-performing model on our validation data were compared to bGS only in the patients with both bGS and psGS available. We inform the balanced accuracy of the predictions, as well as the sensitivity and specificity of the predictions for each GS class.

Since most previous studies saw interest in presenting their results in two categories (lower/intermediate vs high risk), we also present the results of our best-performing model in this fashion (GGG 1–3 vs GGG 4–5).

The statistical analysis was performed using numpy and scipy.stats in Python.

Results

The 101 patients who met the eligibility criteria were grouped by GS into three categories: GS < 8 (GGG 1–3: 60 [59%] patients), GS equal to 8 (GGG 4: 23 [23%] patients), and GS > 8 (GGG 5: 18 [18%] patients). In each cross-validation cycle, two-thirds of the patients (n = 67) were selected from each class as training data (GGG 1–3: 40, GGG 4: 15, GGG 5: 12), leaving the rest (n = 34) as validation (20, 8, and 6 patients, respectively). The best-performing model (highest average bAcc in the validation) for each image modality was selected for comparison. The characteristics of the selected models are summarized in the Supplementary Table 3.

Three out of four single-modality models (PET, T1w, T2w) used a combination of first order, shape, and textural radiomics features, except the ADC model, which did not include any first order features. All double-modality models (PET + T1w, PET + T2w, PET + ADC) included both PET and MR features.

Across the PET-only model and all double-modality models (all of which include PET radiomics), the most often selected PET features were the shape feature "Maximum 2D Diameter Row" (present in all 4 models) and the quantitative biomarker "Total SUV" (present in 3 out of 4 models).

In both models including ADC features, textural ADC features were predominant over shape and first order features (with ratios 6:1 and 5:2 textural features over the rest, respectively). In both models including T2w features, shape features were the most frequent (4:3 and 5:1, respectively). In the models with T1w features, the selected feature ratio was more evenly distributed among the 3 types (first order/shape/textural: 2:1:6 and 3:1:3, respectively).

The performances of the implemented models are summarized in Table 2. For a visual comparison, Fig. 4 displays a box plot of the bAcc of all radiomic and baseline models.

The comparisons between models were performed using the validation bAcc (mean ± standard deviation). For the patient-data baseline model and the radiomics baseline model, the bAcc was 58 ± 5% and 65 ± 7%, respectively. All single-modality models (T1w: 72 ± 3%; T2w: 73 ± 2%; ADC: 76 ± 6%, PET: 75 ± 5%) provided a significantly better classification performance than the patient-data baseline ($p < 0.001$). The radiomics baseline model, while exhibiting a higher bAcc, did not significantly outperform the patient-data baseline. Most single-modality models outperformed also the radiomics baseline model, except the T1w model (T2w: $p = 0.034$, ADC: $p = 0.018$, PET: $p = 0.018$, T1w model: $p = 0.060$). Among the single-modality models, the model trained with ADC radiomics provided the highest performance (76 ± 6%), although not statistically higher than any of the

Table 2 Performances of the trained models (*top*, *grey*: baseline models; *center*, *green*: single-image radiomics; *bottom*, *orange*: double-image radiomics) on the validation dataset, expressed as their balanced accuracies, sensitivities, and specificities (mean and standard deviation, in percentages)

Model	Balanced accuracy [%]		Sensitivity [%]						Specificity [%]					
			GGG 1-3		GGG 4		GGG 5		GGG 1-3		GGG 4		GGG 5	
	mean	stdv	mean	stdv	mean	stdv	mean	stdv	mean	stdv	mean	stdv	mean	stdv
patient baseline	57.55	4.98	63.48	21.35	45.00	24.32	64.17	30.06	74.44	11.26	78.10	14.01	84.37	6.51
radiomics baseline	65.11	7.17	74.51	11.09	50.00	8.16	70.83	23.17	74.03	13.08	82.07	10.32	91.59	7.80
T1w	71.74	2.69	81.89	9.99	58.33	19.54	75.00	14.43	67.86	2.06	92.08	10.23	92.08	5.15
T2w	72.60	2.20	80.29	6.98	72.92	18.04	64.58	26.60	74.70	7.01	89.15	9.29	92.06	7.36
ADC	76.12	6.22	78.37	12.71	79.17	21.65	70.83	29.76	82.59	12.00	84.53	10.00	95.20	5.48
PET	74.90	4.53	72.55	6.63	73.81	15.97	78.33	6.87	86.06	6.35	83.99	4.72	90.11	8.19
PET+T1w	73.85	4.90	75.23	11.32	69.64	9.43	76.67	15.28	78.98	12.31	85.99	6.41	91.52	7.92
PET+T2w	70.96	4.09	74.77	13.44	67.26	22.98	70.83	18.16	75.80	10.00	83.19	9.64	93.49	3.91
PET+ADC	81.57	5.24	70.91	9.58	82.14	15.57	91.67	8.33	91.99	8.03	83.87	7.27	90.22	4.53

other single-modality radiomics models ($p > 0.05$). The sensitivities of this model were similar across classes, but slightly higher for GGG 5 (sens: GGG 1–3: 73 ± 7%; GGG 4: 74 ± 16%; GGG 5: 78 ± 7).

A model trained with combined features from PSMA-PET and ADC map radiomics yielded the highest overall accuracy (82 ± 5%). It significantly outperformed most single-modality models, including the PET-only (with $p = 0.042$), T1w-only ($p = 0.002$), and T2w-only ($p = 0.003$), while the difference with the best model trained with ADC-only radiomics was not statistically significant (ADC: 76 ± 6%, $p = 0.138$). The better performance of this model resulted from the highest sensitivity to the higher risk groups but at the cost of a lower sensitivity to the low/intermediate-risk group (sens: GGG 1–3: 71 ± 10%; GGG 4: 82 ± 16%; 92 ± 8%).

The addition of T1w or T2w radiomics to PSMA-PET radiomics (74 ± 5% and 71 ± 4%, respectively) did not significantly alter the performance of single-modality models ($p > 0.05$). Combined PET + ADC radiomics significantly outperformed both of these double-modality models (PET + ADC: 82 ± 5%; p-value: 0.026 and 0.003, respectively).

P-values for all model combinations are summarized in Table 3.

For the best-performing model, the results of the classification of lower/intermediate- vs high-risk Gleason categories (GGG 1–3 vs GGG 4–5) are presented in Table 4. The best-performing baseline model was again the radiomics baseline (74 ± 5%), although not significantly better than the patient baseline (69 ± 8%, $p = 0.203$). The hybrid PET + ADC model (82 ± 6%) outperformed the patient and radiomics baseline models overall ($p = 0.011$ and 0.029, respectively), with a much higher sensitivity to GGG 4–5 (patient baseline: 74 ± 11; radiomics baseline: 74 ± 13; PET + ADC: 94% ± 8) at the cost of a slightly poorer sensitivity to GGG 1–3 (patient baseline: 63 ± 21%; radiomics baseline: 75 ± 11%; PET + ADC: 71 ± 10).

From the 71 patients with both available psGS and bGS (GGG 1–3: 62% (n = 44); GGG 4: 20% (n = 14); GGG 5: 18% (n = 13)), our model outperformed the bGS in predicting psGS overall (bAcc: 82.5% vs 72.4%, respectively)

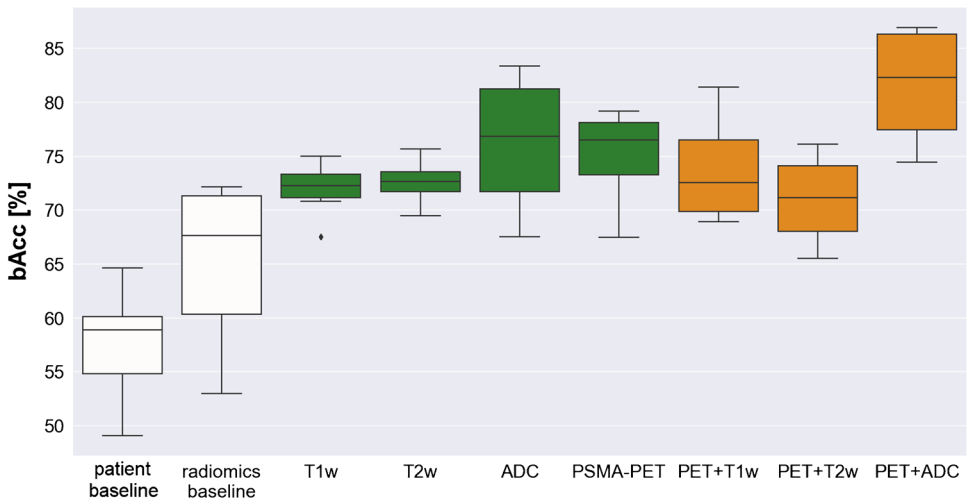


Fig. 4 Boxplot of the balanced accuracies of all best-performing models on the validation sets

Table 3 Correlation matrix (p-values) from unpaired t-tests between all model performances (balanced accuracies) and Shapiro–Wilk normality test significance.

T-test								
p-value	patient baseline	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
	radiomics baseline	0.060	0.034*	0.018*	0.018*	0.033*	0.113	<0.001***
	T1w		0.558	0.144	0.173	0.377	0.705	0.002**
	T2w			0.221	0.289	0.581	0.407	0.003**
	ADC				0.706	0.499	0.120	0.138
	PET					0.708	0.145	0.042*
	PET+T1w						0.448	0.026*
	PET+T2w							0.003**
	Model	T1w	T2w	ADC	PET	PET +T1w	PET +T2w	PET +ADC

Shapiro-Wilk normality									
Model	patient baseline	radiomics baseline	T1w	T2w	ADC	PET	PET+T1w	PET+T2w	PET+ADC
p	0.813	0.303	0.442	0.780	0.602	0.338	0.448	0.780	0.363

Significance levels: *p < 0.05; **p < 0.01; ***p < 0.001; otherwise: not statistically significant (*grey*: baseline models; *green*: single-image models; *orange*: double-image models)

and within each group (sensitivity: GGG 1–3: 72.3% vs 68.2%; GGG 4: 84.0% vs 64.3%; GGG 5: 91.3% vs 84.6%, respectively). The comparisons between biopsy GS and the PET + ADC model in predicting psGS can be found in Table 5.

Discussion

In this work, we studied the potential of hand-crafted radiomics from [^{68} Ga]Ga-PSMA-11 PET/MR images for predicting postsurgical Gleason scores. Previous to this study, the literature on the impact of PET/MRI on PCa radiomics was scarce. In addition, we used a prediction of three GS categories, instead of two as in most previous studies. Important aspects of our work are discussed below: the rationale in choosing our segmentation methods, the results of the feature selection and GS predictions, and limitations and perspectives on future work.

Our segmentation strategy implied a trade-off, which intended to minimize the complexity while achieving effective radiomics. In VOIs with a small number of voxels, radiomics analysis cannot provide much complementary information to the VOI volume [23]. Since PET images have large and rather low-resolution voxels, small prostate lesions are prone to render many radiomic features into surrogates of the number of voxels (i.e., lesion volume). To overcome this limitation, we segmented the whole-prostate gland. Overall, image-based models performed significantly better than the radiomic baseline, indicating that the classification was not merely based on volume surrogates.

Another argument for whole-prostate segmentations was to reduce the complexity of the radiomics workflow. In the case of multi-lesion PCa, a feature extraction process using multiple VOIs would have implied a more complex mechanism to aggregate the radiomics features throughout the lesions. To make matters more complex, PCa lesions are not always simultaneously present in all image modalities, making the segmentations and the feature extraction more cumbersome or even impossible for different images.

Table 4 Performances of the baselines vs the best-performing model trained on 3 GS groups (GGG 1–3, GGG 4, GGG 5) and tested on 2 GS groups (GGG 1–3, GGG 4–5) (*top*, *grey*: baseline models; *bottom*, *orange*: double-image radiomics), expressed as their balanced accuracies (mean and standard deviation, in percentages), and sensitivities by class (t-test comparisons: patient baseline vs radiomics baseline: p = 0.203; patient baseline vs PET + ADC: p = 0.011; radiomics baseline vs PET + ADC: 0.029)

Model	Balanced accuracy [%]		Sensitivity GGG 1-3 [%]		Sensitivity GGG 4-5 [%]	
	mean	stdv	mean	stdv	mean	stdv
patient baseline	68.96	8.39	63.48	21.35	74.44	11.26
radiomics baseline	74.27	4.56	74.51	11.09	74.03	13.07
PET+ADC	82.45	6.40	70.91	9.58	93.99	8.03

Table 5 Comparison of the prediction of postsurgical GS (psGS) between our best-performing model (*orange*: PET + ADC radiomics) and the biopsy GS (bGS, *white*) on the patients with both psGS and bGS available, expressed as their balanced accuracies, sensitivities, and specificities (in percentages)

Model	Balanced accuracy [%]	Sensitivity [%]			Specificity [%]		
		GGG 1-3	GGG 4	GGG 5	GGG 1-3	GGG 4	GGG 5
PET+ADC	82.53	72.29	84.00	91.30	91.66	83.02	93.52
Biopsy GS	72.36	68.18	64.29	84.62	96.30	82.46	82.76

Our whole-prostate approach ensured not only big enough VOIs (of more than one thousand voxels) but also a simpler segmentation and feature extraction process, with only one VOI per image type. This is also a logical approach from a clinical perspective, since one patient would have different GS values across lesions, but only the highest (index lesion) is considered for treatment and prognosis.

After extracting the features from different VOIs, the weakest features were removed through recursive feature elimination (RFE). RFE works by fitting an SVM model with all the features, and eliminating the feature that the SVM considers less relevant. This process is repeated several times, eliminating one feature at a time, so as to leave only the highest ranked and less interdependent ones. One characteristic of all models was that the prostate radius was always selected as a feature, which may be accounting for the effect of prostate volume in the detection of prostate cancer lesions [17]. In particular, the models containing T2w radiomic features relied mostly on this and other shape features, relegating first order and textural features to a lesser role. These selected radiomic features not only reflect image information that is important for the classification but are also influenced by the chosen segmentation method [31]. The models with ADC features, on the other hand, were the two best-performing models and relied mainly on the textural properties of the prostate, instead of its shape features. Interestingly, most PET models repeatedly selected a commonly used quantitative biomarker based on PSMA-TL, stressing its importance in predicting GS.

The best-performing single-modality models included PSMA-PET (bAcc = 75 ± 5%) and ADC map (bAcc = 76 ± 6%) features, which concordantly are the focus of most previous literature [16–20]. Overall, the best-performing model was the one that combined features from both highest-performing single modalities (PET + ADC, bAcc = 82 ± 5%), demonstrating the added value of the multi-modality approach compared to either PET-only or MR-only radiomics. It is interesting to note that the better performance of this model emerged from a higher sensitivity to high-risk groups (GGG 4 and 5), trading sensitivity in the prediction of the low/intermediate-risk groups (GGG 1–3), which implies a misdiagnose (overgrading) of a considerable percentage of low/intermediate-risk patients. As a trade-off, the specificity of low/intermediate-risk patients and the sensitivity to high-risk patients are greatly improved.

All image-based models performed significantly better than the patient-data baseline, demonstrating that image radiomics provide additional information to the available clinical parameters. Most models also significantly outperformed the radiomics baseline, except for the T1w and PET + T2w models. This implies that, for most models, the selected combinations of features were not mere surrogates of volume or intensity, even though some features that correlate with volume were involved in the classification.

The results from the predictions of low/intermediate- vs high-risk Gleason scores require a special discussion. First, it is important to clarify that, although our model allow the predictions of two categories (by combining the predictions of both upper categories, GGG 4 and GGG 5), it was trained for the prediction of three categories. Since it was not optimally trained for this prediction, it would most likely underperform in comparison to a real two-category model. Second, we trained our model to optimize the balanced accuracy, which has no bias towards a particular category. In combining two categories, we are imposing a bias towards these categories, hindering the remaining category. Even so, the overall predictions of the PET + ADC model (82 ± 6%) were still superior to both baseline models (69 ± 8% and 74 ± 5%), at the cost of a lower sensitivity to the low/intermediate-risk category.

As we mentioned, our study has potential clinical implications. Given that not every patient undergoes radical prostatectomy, the biopsy GS is usually used instead of the postsurgical GS in the clinical routine, even though they are not always equivalent. Our hybrid radiomics model outperformed the use of biopsy GS in estimating the postsurgical GS (82.5% vs 72.4%). It is important to note that our dataset lacked biopsy GS for around 30% of the patients. With a larger initial cohort (implying more patients with simultaneous postsurgical and biopsy GS data), we would be able to analyze the power of combined biopsy GS and image radiomics in the prediction of postsurgical GS. As an alternative, we could also use the biopsy GS as part of the patient-data baseline, as a more clinically relevant baseline model.

Our study is not without limitations. Firstly, family-wise error rates (FEWRs) across the statistical analyses were not controlled, meaning that several comparisons between models were performed through a statistical test (i.e., Student's t-test), without correcting for the higher probability of Type I errors. Secondly, our results show

high intragroup variances in the performance of most classifiers (Table 2). In particular, even though the model trained with PET + ADC radiomics outperformed the rest of the models in terms of balanced accuracy, the difference with respect to the ADC-only model was not statistically significant. The high variances can be partly attributed to the small number of patients and strongly unbalanced dataset, which implies that, in each fold, only radiomics from around 15 or 12 patients (GGG 4 or GGG 5, respectively) were used for training the models, and 8 or 6 patients (GGG 4 or GGG 5, respectively) were used for evaluation in the less prevalent GS classes. More robust models with lower variability between successive training cycles would require a larger cohort and, ideally [34], an external testing cohort.

Although GS is widely used as a proxy for the aggressiveness of PCa, it must be used with caution. Studies show that, in some cases, the fraction of Gleason patterns relate to the outcome of patients better than the GS [35]. For instance, GS 7 tumors represent a rather diverse population and, as a consequence, there is clinical value in further differentiating it in subcategories. In our work, we focused on demonstrating that [^{68} Ga]Ga-PSMA-11 PET and mpMRI would synergically work in the prediction of GS, but our results could profit from considering the outcome of the patients beyond GS.

In our work, we took advantage of the benefits of whole-prostate segmentations, but our segmentation approach has room for improvement. The PI-RADS v2 protocol [36] proposes a two-region approach for PCa diagnosis with mpMRI. According to the protocol, DWI/ADC map is the most informative sequence for the assessment of PCa in the peripheral zone (PZ), while T2w is used mainly for assessment of PCa in the transition zone (TZ) of the prostate. To implement this, PZ/TZ segmentation would mean a more demanding segmentation work for the radiologists, although, for big cohorts, it could also be automated by appropriately training a deep learning (DL)-based algorithm.

DL techniques can also be exploited for feature extraction and prediction. Our decision of using hand-crafted radiomics was based on their higher level of interpretability, as well as on the existence of several previous studies. On the contrary, DL features are more complex to decipher, requiring specific methodologies to "open the black box" and provide an explanation of the output classification. One such methodology is the use of activation or saliency maps of attention, and their relation to correct and incorrect classifications [37, 38]. Another applicable approach is training a fully convolutional network for segmentation of a pathological tissue (e.g., a tumor), and using its trained semantic layers as features for the classification of the pathology [39]. An explainable DL approach could help save time and impact the performance of our predictions, especially in the case of bigger patient cohorts. There are already several works predicting GS from mpMRI-only radiomics using non-explainable deep convolutional layers [40–42]. Based on our results and available technology, a step forward would benefit from including also PSMA-PET radiomics and some form of explainable DL approach.

Conclusion

Our work shows promising results on the combined power of PSMA-PET and mpMRI radiomic features for predicting postsurgical GS in PCa patients and envisions a reliable tool that helps urologists and radiologists in their daily decision-making process.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-021-05430-z.

Funding Open Access funding enabled and organized by Projekt DEAL. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764458.

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

Code availability Questions regarding the code implemented in this work can be addressed to the corresponding author.

Declarations

Ethics approval The retrospective data analysis was approved by the medical ethics committee of the Technical University of Munich (reference number: 5665/13S).

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. https://doi.org/10.3322/caac.21492.
2. Maurer T, Eiber M, Schwaiger M, Gschwend JE. Current use of PSMA-PET in prostate cancer management. Nat Rev Urol. 2016;13(4):226–35. https://doi.org/10.1038/nrurol.2016.26.
3. Sciarra A, Barentsz J, Bjartell A, Eastham J, Hricak H, Panebianco V, Witjes JA. Advances in magnetic resonance imaging: how they are changing the management of prostate cancer. Eur Urol. 2011;59(6):962–77. https://doi.org/10.1016/j.eururo.2011.02.034.
4. Fendler WP, Eiber M, Beheshti M, Bomanji J, Ceci F, Cho S, Giesel F, Haberkorn U, Hope TA, Kopka K, Krause BJ, Mottaghy FM, Schöder H, Sunderland J, Wan S, Wester HJ, Fanti S, Herrmann K. Joint EANM and SNMMI procedure guideline for prostate cancer imaging: version 1.0. Eur J Nucl Med Mol Imaging. 2017;44(6):1014–24. https://doi.org/10.1007/s00259-017-3670-z.
5. Maurer T, Gschwend JE, Rauscher I, Souvatzoglou M, Haller B, Weirich G, Wester HJ, Heck M, Kübler H, Beer AJ, Schwaiger M, Eiber M. Diagnostic efficacy of (68)gallium-PSMA positron emission tomography compared to conventional imaging for lymph node staging of 130 consecutive patients with intermediate to high risk prostate cancer. J Urol. 2016;195(5):1436–43. https://doi.org/10.1016/j.juro.
6. Calais J, Ceci F, Eiber M, Hope T, Hofman M, Rischpler C, Bach-Gansmo T, Nanni C, Savir-Baruch B, Elashoff D, Grogan T, Dahlbom M, Slavik R, Gartmann J, Nguyen K, Lok V, Jadvar H, Kishan A, Rettig M, Czernin J. 18F-fluciclovine PET-CT and 68Ga-PSMA-11 PET-CT in patients with early biochemical recurrence after prostatectomy: a prospective, single-centre, single-arm, comparative imaging trial. Lancet Oncol. 2019;20(9):1286–94. https://doi.org/10.1016/S1470-2045(19)30415-2.
7. Treglia G, Pereira Mestre R, Ferrari M, Bosetti DG, Pascale M, Oikonomou E, De Dosso S, Jermini F, Prior JO, Roggero E, Giovanella L. Radiolabelled choline versus PSMA PET/CT in prostate cancer restaging: a meta-analysis. Am J Nucl Med Mol Imaging. 2019;9(2):127–39.
8. Eiber M, Weirich G, Holzapfel K, Souvatzoglou M, Haller B, Rauscher I, Beer AJ, Wester HJ, Gschwend J, Schwaiger M, Maurer T. Simultaneous 68Ga-PSMA HBED-CC PET/MRI improves the localization of primary prostate cancer. Eur Urol. 2016;70(5):829–36. https://doi.org/10.1016/j.eururo.2015.12.053.
9. Maurer T, Gesterkamp H, Nguyen N, Westenfelder K, Gschwend JE, Budäus L, Rauscher I, Vag T, Weber W, Eiber M. "68Ga-PSMA-11 PET/mpMRI for local detection of primary prostate cancer in men with a negative prior biopsy". Aktuelle Urol. 2020. https://doi.org/10.1055/a-1198-2305.
10. Giesel FL, Sterzing F, Schlemmer HP, Holland-Letz T, Mier W, Rius M, Afshar-Oromieh A, Kopka K, Debus J, Haberkorn U, Kratochwil C. Intra-individual comparison of (68)Ga-PSMA-11-PET/CT and multi-parametric MR for imaging of primary prostate cancer. Eur J Nucl Med Mol Imaging. 2016;43(8):1400–6. https://doi.org/10.1007/s00259-016-3346-0.
11. Egevad L, Delahunt B, Srigley JR, Samaratunga H. International Society of Urological Pathology (ISUP) grading of prostate cancer—an ISUP consensus on contemporary grading. APMIS. 2016;124(6):433–5. https://doi.org/10.1111/apm.12533.
12. Cohen MS, Hanley RS, Kurteva T, Ruthazer R, Silverman ML, Sorcini A, Hamawy K, Roth RA, Tuerk I, Libertino JA. Comparing the Gleason prostate biopsy and Gleason prostatectomy grading system: the Lahey Clinic Medical Center experience and an international meta-analysis. Eur Urol. 2008;54(2):371–81. https://doi.org/10.1016/j.eururo.2008.03.049.
13. Valdora F, Houssami N, Rossi F, Calabrese M, Tagliafico AS. Rapid review: radiomics and breast cancer. Breast Cancer Res Treat. 2018;169(2):217–29. https://doi.org/10.1007/s10549-018-4675-4.
14. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, Madabhushi A. Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer. 2018;115:34–41. https://doi.org/10.1016/j.lungcan.2017.10.015.
15. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing the cancer genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 2017;4:170117. https://doi.org/10.1038/sdata.2017.117.
16. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, Sala E, Hricak H, Deasy JO. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci U S A. 2015;112(46):E6265–73. https://doi.org/10.1073/pnas.1505935112.
17. Chaddad A, Kucharczyk MJ, Niazi T. Multimodal radiomic features for the predicting Gleason score of prostate cancer. Cancers (Basel). 2018;10(8):249. https://doi.org/10.3390/cancers10080249.
18. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solórzano G, Erho N, Balagurunathan Y, Punnen S, Davicioni E, Gillies RJ, Pollack A. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res. 2016;5(4):432–47. https://doi.org/10.21037/tcr.2016.06.20.
19. Zamboglou C, Carles M, Fechter T, Kiefer S, Reichel K, Fassbender TF, Bronsert P, Koeber G, Schilling O, Ruf J, Werner M, Jilg CA, Baltas D, Mix M, Grosu AL. Radiomic features from PSMA PET for non-invasive intraprostatic tumor discrimination and characterization in patients with intermediate- and high-risk prostate cancer—a comparison study with histology reference. Theranostics. 2019;9(9):2595–605. https://doi.org/10.7150/thno.32376.
20. Cysouw MCF, Jansen BHE, van de Brug T, Oprea-Lager DE, Pfaehler E, de Vries BM, van Moorselaar RJA, Hoekstra OS, Vis AN, Boellaard R. Machine learning-based analysis of [18-F]-DCFPyL PET radiomics for risk stratification in primary prostate cancer. Eur J Nucl Med Mol Imaging. 2020. https://doi.org/10.1007/s00259-020-04971-z.
21. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, Vickers AJ, Parwani AV, Reuter VE, Fine SW, Eastham JA, Wiklund P, Han M, Reddy CA, Ciezki JP, Nyberg T, Kleine EA. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. Eur Urol. 2016;69(3):428–35. https://doi.org/10.1016/j.eururo.2015.06.046.
22. Martin R, Jüttler S, Müller M, Wester HJ. Cationic eluate pretreatment for automated synthesis of [^{68}Ga]CPCR4.2. Nucl Med Biol. 2014;41(1):84–9. https://doi.org/10.1016/j.nucmedbio.2013.09.002.
23. Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, Hindié E, Martineau A, Pradier O, Hustinx R, Perdrisot R, Guillevin R, El Naqa I, Visvikis D. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. J Nucl Med. 2015;56(1):38–44. https://doi.org/10.2967/jnumed.114.144055.
24. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET." IEEE Trans Med Imaging. 2009;28(6):881–93. https://doi.org/10.1109/TMI.2008.2012036.

25. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, Lambin P, Pradier O, Visvikis D. "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications." Int J Radiat Oncol Biol Phys. 2010;77(1):301–8. https://doi.org/10.1016/j.ijrobp.2009.08.018.
26. Zwanenburg A, Leger S, Vallières M, Löck S. "Image biomarker standardisation initiative," arXiv preprint. 2019. arXiv:1612.07003v11.
27. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
28. Gafita A, Bieth M, Kroenke M, Tetteh G, Guenther E, Menze B, Weber WA, Eiber M. "qPSMA: a semi-automatic software for whole-body tumor burden assessment in prostate cancer using 68Ga-PSMA11 PET/CT". J Nucl Med. 2019. https://doi.org/10.2967/jnumed.118.224055.
29. Schmuck S, von Klot CA, Henkenberens C, Sohns JM, Christiansen H, Wester HJ, Ross TL, Bengel FM, Derlin T. Initial experience with volumetric 68Ga-PSMA I&T PET/CT for assessment of whole-body tumor burden as a quantitative imaging biomarker in patients with prostate cancer. J Nucl Med. 2017;58(12):1962–8. https://doi.org/10.2967/jnumed.117.193581.
30. Schmidkonz C, Cordes M, Schmidt D, Bäuerle T, Goetz TI, Beck M, Prante O, Cavallaro A, Uder M, Wullich B, Goebell P, Kuwert T, Ritt P. 68Ga-PSMA-11 PET/CT-derived metabolic parameters for determination of whole-body tumor burden and treatment response in prostate cancer. Eur J Nucl Med Mol Imaging. 2018;45(11):1862–72. https://doi.org/10.1007/s00259-018-4042-z.
31. Domachevsky L, Bernstine H, Goldberg N, Nidam M, Stern D, Sosna J, Groshar D. Early 68GA-PSMA PET/MRI acquisition: assessment of lesion detectability and PET metrics in patients with prostate cancer undergoing same-day late PET/CT. Clin Radiol. 2017;72(11):944–50. https://doi.org/10.1016/j.crad.2017.06.116.
32. Solari EL, Gafita A, Visvikis D, Weber W, Eiber M, Hatt M, Nekolla SG. "Complementary diagnostic value of PSMA PET and MR radiomics for prostate cancer staging". in Eur J Nucl Med Mol Imaging. 2020;47:1–753. https://doi.org/10.1007/s00259-020-04988-4. European Association of Nuclear Medicine October 22 – 30, 2020 Virtual.
33. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, Huang SH, Purdie TG, O'Sullivan B, Aerts HJWL, Jaffray DA. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol. 2019;130:2–9. https://doi.org/10.1016/j.radonc.2018.10.027.
34. Hatt M, Lucia F, Schick U, Visvikis D. Multicentric validation of radiomics findings: challenges. EBioMedicine (Commentary). 2019;47:20–1. https://doi.org/10.1016/j.ebiom.2019.08.054.
35. Sauter G, Steurer S, Clauditz TS, Krech T, Wittmer C, Lutz F, Lennartz M, Janssen T, Hakimi N, Simon R, von Petersdorff-Campen M, Jacobsen F, von Loga K, Wilczak W, Minner S, Tsourlakis MC, Chirico V, Haese A, Heinzer H, Huland H, Schlomm T. Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. Eur Urol. 2016;69(4):599–600. https://doi.org/10.1016/j.eururo.2015.10.029.
36. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur Urol. 2019;76(3):340–51. https://doi.org/10.1016/j.eururo.2019.02.033.
37. Böhle M, Eitel F, Weygandt M, Ritter K. "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification". Front Aging Neurosci. 2019;11(194). https://doi.org/10.3389/fnagi.2019.00194.
38. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F, Müller KR, Binder A. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Rep. 2020;10:6423. https://doi.org/10.1038/s41598-020-62724-2.
39. Baek S, He Y, Allen BG, Buatti JM, Smith BJ, Tong L, Sun Z, Wu J, Diehn M, Loo BW, Plichta KA, Seyedin SN, Gannon M, Cabel KR, Kim Y, Wu X. "Deep segmentation networks predict survival of non-small cell lung cancer," ArXiv. Image and Video Processing (eess.IV). 2019. abs/1903.11593v2.
40. Brunese L, Mercaldo F, Reginelli A, Santone A. Radiomics for Gleason score detection through deep learning. Sensors (Basel). 2020;20(18):5411. https://doi.org/10.3390/s20185411.
41. Zong W, Lee J, Pantelic M, Wen N. "Prediction of Gleason grade group of prostate cancer on multiparametric MRI using deep machine learning models". Proceedings of the American Radium Society's 102nd Annual Meeting. 2020. https://doi.org/10.1016/j.ijrobp.2020.02.484.
42. Nagpal K, Foote D, Liu Y, Cameron Chen PH, Wulczyn E, Tan F, Olson N, Smith JL, Mohtashamian A, Wren JH, Corrado GS, MacDonald R, Peng LH, Amin MB, Evans AJ, Sangoi AR, Mermel C, Hipp J, Stumpe MC. "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer", npj Digit. Med. 2019;2:48. https://doi.org/10.1038/s41746-019-0112-2.

Authors and Affiliations

Esteban Lucas Solari[1] · Andrei Gafita[1] · Sylvia Schachoff[1] · Borjana Bogdanović[1] · Alberto Villagrán Asiares[1] · Thomas Amiel[2] · Wang Hui[1] · Isabel Rauscher[1] · Dimitris Visvikis[3] · Tobias Maurer[4] · Kristina Schwamborn[5] · Mona Mustafa[1] · Wolfgang Weber[1] · Nassir Navab[6] · Matthias Eiber[1] · Mathieu Hatt[3] · Stephan G. Nekolla[1]

1 School of Medicine, Department of Nuclear Medicine, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

2 School of Medicine, Department of Urology, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

3 INSERM, UMR 1101, LaTIM, Univ Brest, Brest, France

4 Department of Urology and Martini-Klinik Prostate Cancer Center, University Hospital Hamburg-Eppendorf, Hamburg, Germany

5 School of Medicine, Institute of Pathology, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

6 School of Computer Science, Computer Aided Medical Procedures and Augmented Reality, Technical University Munich, Munich, Germany