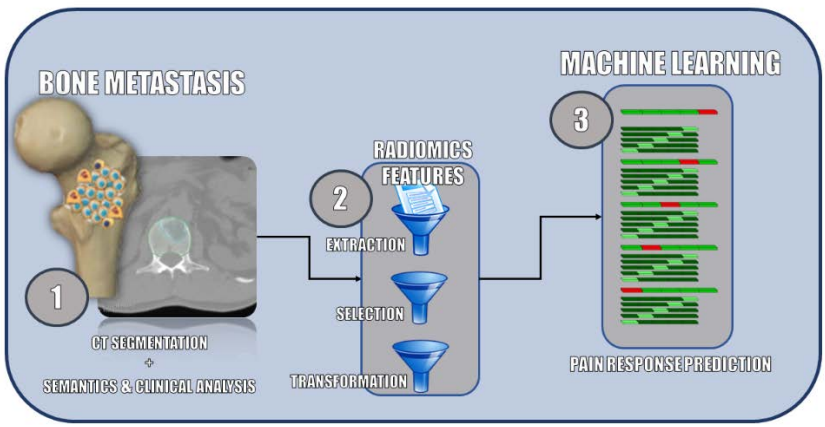
Article

Machine Learning based Prediction of Pain Response to Palliative Radiation Therapy - is there a Role for Planning CT-based Radiomics and Semantic Imaging Features?

Óscar Llorián-Salvador[1,2,3], Joachim Akhgar[1], Steffi Pigorsch[1], Kai Borm[1], Stefan Münch[1], Denise Bernhardt[1,4,5], Burkhard Rost[2], Miguel A. Andrade-Navarro[3], Stephanie E. Combs[1,4,5], Jan C. Peeken[1,4,5]

1 Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany
2 Department for Bioinformatics and Computational Biology, Informatik 12, Technical University of Munich (TUM), Garching, Germany
3 Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany
4 Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Zentrum München, Germany
5 Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich
* Correspondence: Jan C. Peeken, jan.peeken@tum.de; Tel.: +498941404501; Mailing address: Klinik für RadioOnkologie und Strahlentherapie, Universitätsklinikum rechts der Isar, Technische Universität München (TUM), Ismaninger Straße 22, 81675 München, Germany

Abstract: Background: Painful spinal bone metastases (PSBMs) patients regularly receive palliative radiation therapy (RT) with response rates in about 2 of 3 patients. In this exploratory study, we evaluated the value of machine learning (ML) models based on radiomic, semantic and clinical features to predict complete pain response. Methods: Gross tumour volumes (GTV) and clinical target volumes (CTV) of 261 PSBMs were segmented on planning computed tomography (CT) scans. Radiomic, semantic and clinical features were collected for all patients. Random forest (RFC) and support vector machine (SVM) classifiers were compared using repeated nested cross-validation.Results: The best radiomic classifier was trained on CTV with an area under the receiver-operator curve (AUROC) of 0.62 ± 0.01 (RFC; 95% confidence interval). The semantic model achieved a comparable AUROC of 0.63 ± 0.01 (RFC), significantly below the clinical model (SVM, AUROC: 0.80 ± 0.01); and slightly lower than the spinal instability neoplastic score (SINS; LR, AUROC: 0.65 ± 0.01). A combined model did not improve performance (AUROC: 0,74 ± 0,01). Conclusions: We could demonstrate that radiomic and semantic analyses of planning CTs allowed for limited prediction of therapy response to palliative RT. ML predictions based on established clinical parameters achieved the best results.



Keywords: radiomics; machine learning; radiation therapy; bone metastases; prediction

1. Introduction

Different techniques have been used over the years to find early signs of cancer and its evolution, such as screening programs. machine learning (ML) techniques, as an innovative way of studying this disease, appeared approximately 30 years ago [1–3].

Machine learning is a field of computer science that uses statistical and probabilistic models, helped by optimization techniques, to analyse past examples (training set) and extract valuable information from it. This information is learned through the detection of patterns in the known examples of the data, and used to predict the most likely outcome of new cases (test set / final application) [4,5].

There is a significant amount of cancer research based on ML techniques, applying different ML algorithms such as support vector machines (SVMs) and random forest classifiers (RFCs) [6–8]. One field that has experienced a rapid growth over the last few years thanks to the use of ML techniques to extract information from these features is radiomics [9,10].

Radiomics is a medical field focusing on the extraction of quantitative features from medical imaging (i.e., computer tomography (CT)) [11,12]. Radiomics information is used for training machine learning models to predict clinical or biological endpoints [13,14]. Radiomics has been applied to various cancer types for prediction of survival, disease progression, tumour response, molecular aberrations and detection of metastases or areas of infiltrative tumour [15–26]. However, applications of radiomic analysis specific for the prediction of non-tumour response to radiotherapy (RT) have not been thoroughly studied. Some studies have investigated the prediction of RT-dependent side effects such as xerostomia or pneumonitis [27,28].

Painful spinal bone metastases (PSBMs) are regularly treated by palliative RT. About two thirds of the patients experience a partial or complete response in terms of pain reduction [29]. Clinical parameters, however, show limited predictive capabilities to identify patients that profit from palliative RT [30]. The Spinal instability neoplastic score (SINS) has been developed by the Spine Oncology Study Group to assess instability of spinal bone metastases [31]. At the same time, the SINS provides a semantic tool to predict pain response to RT [29].

In this retrospective study we sought to determine the potential of ML-based prediction of RT therapy response of PSBM. Besides clinical features, we investigated whether CT-based radiomic features and semantic features can be used to predict pain response, as well. The best strategy for the definition of volumes of interest (VOI) in regard to macroscopic or microscopic metastatic expansion was assessed for radiomic feature extraction. During modelling, multiple distinct machine learning strategies were developed and compared.

2. Materials and Methods

2.1. Clinical data curation

Patient records of all (n = 491) patients treated with palliative RT for bone metastases between 2009 and 2017 at our institution were analysed. Patients with non-spinal metastases, previous interventions (e.g., surgical stabilization or kyphoplasty) or RT, haematological bone manifestations, and missing information regarding pain response were excluded (Figure S1 for a patient workflow). Patient demographics were assessed for each patient (Table 1 for characteristics of patients, RT and metastatic disease). Clinical parameters previously shown to be associated with pain response such as Karnofsky performance score (KPS), age, use of opioids, and histology (breast cancer, non-small cell lung

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2022 doi:10.20944/preprints202212.0195.v1

cancer (NSCLC) and others) were determined and used as input for the clinical ML models (Table S1 for the exact distribution of histologies) [29,30,32,33]. Histology, as the only categorical value present in the clinical data, was encoded into three dummy binary features. Pain response was rated retrospectively on the basis of patient records following the "international consensus on palliative radiotherapy endpoints for future clinical trials in bone metastases" at the first follow-up visit 6 weeks after RT [34]: complete response: "pain score of 0 at treated site with no concomitant increase in analgesic intake", partial response: "Pain reduction of at least 2 at the treated site (scale of 0 to 10) without analgesic increase, or analgesic reduction of at least 25% without pain increase", pain progression: "increase in pain score of at least 2 or increase of analgesics of at least 25%", and indeterminate response or "no response": "no response or any response not captured by the other categories". The SINS was determined by visual assessment of planning CTs following the definition of the Spine Oncology Study Group [31].

2.2. Definition of VOIs

For each metastasis, two separate VOI definitions were segmented on the planning CT scans using Eclipse 13.0 (Varian Medical Systems, Palo Alto, USA) (Table S2 for acquisition parameters). First, the visible blastic and / or lytic gross tumour volume (GTV) including any adjacent soft-tissue component was manually segmented. Secondly, a clinical target volume (CTV) considering potential microscopic spread was segmented following the International Spine Radiosurgery Consortium Consensus Guidelines for Target Volume Definition in Spinal Stereotactic Radiosurgery [35].

2.3. Radiomics feature extraction

Pre-processing and radiomics feature extraction were performed using the pyRadiomics library (version 2.0) in Python (version 3.6.4) (Figure 1 for study workflow) [36]. For pre-processing, a fixed bin width of 20 was used for image discretization. Isotropic resampling was performed to a voxel size of 1x1x1 mm using Bspline interpolation. 105 radiomics features, including shape, first-order, and texture features were computed from the original image. All extracted features were computed according to the "image biomarker standardization initiative" guidelines (Table S3) [37].

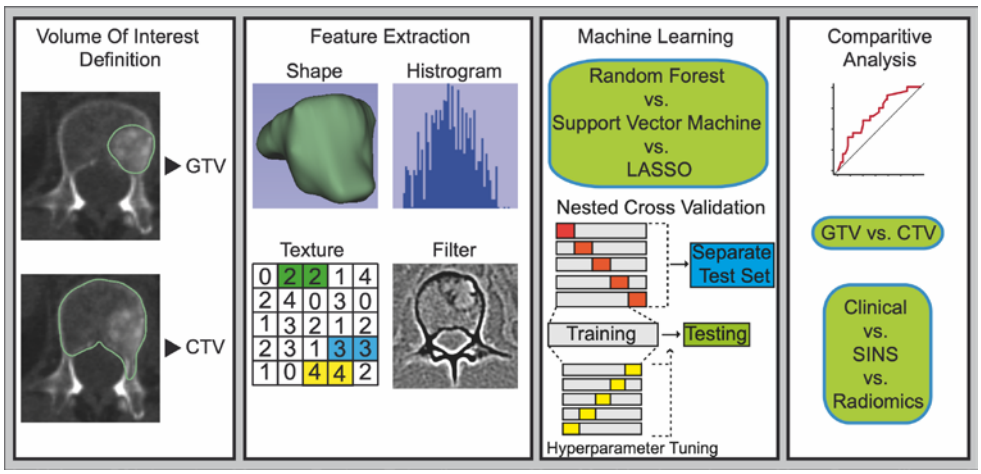


Figure 1. Workflow.

2.4. Semantic feature extraction

Semantic features from the SINS score and other imaging descriptors were determined by an MD student (JA) and controlled by a radiation oncology resident with 3 years of experience (JCP) (Table 1 for a complete listing).

Table 1. List of semantic features.

 doi:10.20944/preprints202212.0195.v1

Feature	Possible values
Imaging - Bone reaction	Blastic reaction
	Mixed reaction (lytic / blastic)
	Lytic reaction
Soft tissue component	Yes
	No
GTV classification	Any portion of vertebral body
	Lateralized within body
	Diffuse within body
	Body + unilateral pedicle
	Body + bilateral pedicle / transverse processBody
	Unilateral pedicle
	Unilateral lamina
	Spinous process
Posterolateral involvement of the spinal elements	Bilateral
	Unilateral
	None of the above
Vertebral body collapse	>50% collapse
	<50% collapse
	No collapse with >50% body involved
	None of the above
Location	Junctional
	Mobile
	Semirigid
	Rigid

2.5. Machine Learning modelling

The number of patients was filtered by removing incomplete entries and taking the intersection of patients with all CTV, GTV, semantic, clinical and SINS data. This resulted in a dataset with 233 pre-processed PSBM with known outcome. For feature reduction, both redundancy reduction and feature correlation to the prediction target were taken into consideration with the Maximum Relevance – Minimum Redundancy (MRMR) algorithm (mrmr-selection library, version 0.2.2) [38]. The best 15 features were selected to be used by the ML algorithms.

Given the small dataset size and to ensure a correct hyperparameter optimization, nested 5-fold cross-validation was applied to train and validate the ML models. The cross-validation splits were stratified by patient ID to evenly distribute multiple samples from the same patient between training, validation and test sets. In order to correct the heavy class imbalance (negative to positive ratio of 10:1), Synthetic Minority Oversampling Technique (SMOTE) was used (imbalanced-learn library, version 0.8.0). To avoid overfitting by class repetition, random minority class oversampling was complemented with random majority class undersampling. The normalisation, feature selection and class imbalance correction steps were performed in the inner fold of the nested cross-validation to avoid data leakage and bias. Nested cross-validation was repeated for 50 iterations, for a total of 250 aggregated models, to increase the statistical strength of the results.

Hyperparameter optimization was performed via exhaustive grid search in the inner fold of the nested cross-validation, using balanced accuracy (BA) as the optimization criteria. SVM and RFC were used for general modelling; and Logistic Regression (LR) was used for the analysis of SINS. All models come from the scikit-learn library [39] (version 0.24.2). Firstly, these models were trained on both segmentation modes: CTV and GTV to assess their predictive quality against a binary prediction target (Table 3): complete pain

 doi:10.20944/preprints202212.0195.v1

response (complete response vs partial response / indeterminate response / no response / pain progression). Results were compared to determine the best modelling strategy. The best model was then compared to clinical, SINS, and semantic models (Table 4). Finally, multiple combined models were devised to assess whether combined models performed better (Tables 5 and S6).

The importance given by models to their features was recorded in order to analyse the feature importance for all models developed. Since it is not possible to track the weight of features for non-linear kernels in SVM, only the percentage of feature selection was shown. For RFC models, this importance is shown as the Gini Importance or mean decrease in impurity of the nodes (the higher, the more important).

2.6. Statistical analysis

Given the small dataset size and, therefore, unclear class distribution, Min-max normalisation was performed to scale all features (scikit-learn library, version 0.24.2), while retaining the same distribution. Outlier detection is performed before the nested cross-validation to avoid extreme values from affecting the distribution of the data (scikit-learn library, version 0.24.2). All error margins are reported as standard errors with a coefficient of 1.96 for a confidence interval covering 95% of the observations. All models were evaluated, principally, using the Area Under the Receiver-Operator Curve (AUROC). In addition, BA, F1 score and Matthews Correlation Coefficient (MCC) were secondarily examined. Statistical analysis and radiomic model building were performed using Python (version 3.7).

3. Results

3.1. Pain response to RT

A retrospective cohort of 90 patients with a total of 267 PSBM fitted the inclusion and exclusion criteria in our institution (Figure S1 for a patient workflow). Mammary carcinoma, prostate carcinoma and NSCLC were the three most frequent (63%) cancer types (Table 2 for patient characteristics and Table S1 for a distribution of all cancer histologies). There was a median of two PSBMs per patient with a total of 41 solitary PSBMs. Partial and complete pain response retrospectively assessed from patient files was achieved in 33% and 52% of patients, respectively.

Table 2. Characteristics of patients, radiotherapy and metastatic disease with complete information.

Patient characteristic			
	Complete Response (n = 30 p)	**Partial or No Response (n = 60 p)**	**p-value[a]**
Gender: Male	14 p (43%)	29 p (48%)	0.82
Gender: Female	16 p (57%)	31 p (52%)	
Age	m 66 (r 26 – 88)	m 66 (r 30 - 87)	0.74
Karnofsky Performance Score	m 70 (r 60 - 100)	m 80 (r 30 - 90)	0.34
Opioid medication	16 p (57%)	37 p (62%)	0.50
Tumour Type	Mammary / prostate carcinoma: 11 p (37%) NSCLC: 7 p (23%) Others: 12 p (40%)	Mammary / prostate carcinoma: 31 p (52%) NSCLC: 8 p (13%) Others: 21 p (35%)	0.30
Partial response	-	47 p (78%)	-
Overall Survival	m 5.5 months (r 0.7 – 55.8 months)	m 7.5 months (r 0.1 – 68.1 months)	0.90

 doi:10.20944/preprints202212.0195.v1

Radiotherapy			
Single dose	m 3 (r 2 - 8)	m 3 (r 2- 8)	0.54
Total dose	m 33 (r 8 - 44)	m 30 (r 8 - 45)	0.10
Number of fractions	m 10 (r 1 -22)	m 10 (r 1- 19)	0.45
Bone Metastases			
Number of metastases	65	196	
Number of metastases per patient	m 1.5 (r 1- 6)	m 2.5 (r 1 - 10)	0.055
Previous RT	0 p (0%)	0 p (0%)	-
Localization	Sacrum: 8 p (12%) Lumbar: 34 p (52%) Thoracic: 21 p (32%) Cervical: 2 p (3%)	Sacrum: 17 p (9%) Lumbar: 83 p (42%) Thoracic: 83 p (42%) Cervical: 13 p (7%)	0.26
Bone reaction	Blastic: 15 p (23%) Lytic: 31p (48%) Mixed: 19 p (29%)	Blastic: 56 p (29%) Lytic: 28 p (14%) Mixed: 112 p (57%)	0.03
Soft tissue component	25 p (38%)	48 p (25%)	0.08
Extent of metastasis[b]	vertebral body: 17 p (26%) body / pedicle: 4 p (6%) body / pedicle / transverse process: 2 p (3%) Unilateral pedicle: 23 p (35%) Unilateral lamina: 18 p (28%) Spinous process: 1 p (2%)	vertebral body: 54 p (28%) body / pedicle: 9 p (5%) body / pedicle / transverse process: 8 p (4%) Unilateral pedicle: 35 p (18%) Unilateral lamina: 88 p (45%) Spinous process: 3 p (2%)	0.03
SINS	m 7 (3 – 14)	m 8 (0 – 15)	0.02

Abbreviation: m: median, p: patients, r: range, SINS: Spinal Instability Neoplastic Score.
[a]Wilcoxon rank sum test for continuous and ordinal variables, Fisher's exact test for nominal variables, log rank test for comparison of survival times. Not corrected for multiple testing.
[b]Following the Gross Tumour Volume (GTV) classification of the International Spine Radiosurgery Consortium [1].

3.2. Determination of the best VOI for radiomic analysis and modelling strategy

Table 3. AUROC, BA, F1 Score and MCC for the best modelling algorithms trained on both radiomic segmentation modes (GTV and CTV).

Segmentation	Model	AUROC	BA	F1	MCC
GTV	SVM	0.58 ± 0.01	0.54 ± 0.02	0.33 ± 0.03	0.08 ± 0.04
CTV	RFC	0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04

The best performing model was a RFC trained on the CTV radiomic segmentation, with the highest overall scores (AUROC: 0.62 ± 0.01) (Table 3 for outcome metrics and Figure 2 for ROC and calibration curves). While the RFC reached the highest performance, the SVM results were more consistent. The best segmentation mode was CTV, with higher performance regardless of the modelling strategy.

 doi:10.20944/preprints202212.0195.v1

3.3. Comparison to clinical baseline, semantic and SINS models

The best segmentation mode among the radiomics models (CTV) was then compared to the clinical, the semantic and the SINS models (Table 4 and Figure 2).

Table 4. AUROC, BA, F1 Score and MCC for the best models, comparing the best radiomics model to the semantic features, clinical baseline and SINS variable.

Data	Model	AUROC	BA	F1	MCC
CTV	RFC	0.62 ± 0.01	0.58 ± 0.02	0.37 ± 0.03	0.15 ± 0.04
Semantic	RFC	0.63 ± 0.01	0.58 ± 0.02	0.39 ± 0.03	0.16 ± 0.04
Clinical	SVM	0.80 ± 0.01	0.72 ± 0.03	0.56 ± 0.05	0.43 ± 0.06
SINS	LR	0.65 ± 0.01	0.58 ± 0.03	0.36 ± 0.05	0.16 ± 0.06
SINS (binary)		0.54 ± 0.01	0.52 ± 0.03	0.19 ± 0.05	0.04 ± 0.06

The semantic features, on the other hand, achieved almost identical results to the best radiomic segmentation: none performed statistically better. Lastly, Logistic Regression (LR) trained only on the SINS variable achieved very different results: SINS (binarized) performed very close to random, with a poor classification quality (MCC: 0.04 ± 0.06); on the other hand, the non-binarized SINS model performed similar to the CTV-based radiomics segmentation model but higher AUROC (0.65 ± 0.01).

The clinical ML model outperformed all other models regardless of the modelling algorithm with statistical significance (Table S5 and Figure S2). The best clinical model (SVM) predicted pain response with a BA of 0.72 ± 0.03 and an AUROC of 0.80 ± 0.01. Given the limited features that the SINS and clinical datasets comprised, their respective prediction models showed a wider standard error on scores where greater variance was expected (F1 and MCC).

3.4. Benefits by combining imaging and clinical features

The SVM was evaluated on the possible performance increase by combining the best radiomics model (CTV), clinical, SINS, and semantic features (Table 4, Figure 2 and Figure S2).

Table 5. AUROC, BA, F1 Score and MCC for SVM models trained on the combination of radiomic, clinical, SINS and semantic features.

Data	Model	AUROC	BA	F1	MCC
CTV + SINS	SVM	0,61 ± 0,01	0,57 ± 0,02	0,36 ± 0,04	0,13 ± 0,04
CTV + Clinical		0,75 ± 0,01	0,69 ± 0,02	0,52 ± 0,03	0,35 ± 0,04
Semantic + SINS		0,62 ± 0,01	0,58 ± 0,02	0,39 ± 0,03	0,15 ± 0,04
Semantic + Clinical		0,68 ± 0,01	0,63 ± 0,02	0,45 ± 0,03	0,24 ± 0,04
CTV + SINS + Clinical + Semantic		0,67 ± 0,01	0,62 ± 0,02	0,44 ± 0,03	0,22 ± 0,04

The best performance with combined models was achieved with a SVM trained on CTV and clinical data (AUROC: 0.75 ± 0.01). The addition of non-binarized SINS did not significantly affect the performance of any combined model. An SVM model trained on all data (CTV, non-binarized SINS, clinical and semantic features) outperformed one using only radiomic data; however, it was significantly worse than the best combined model. Interestingly, a model trained only on semantic and clinical features achieved the same performance level as the combined model with all available features (AUROC: 0.68 ± 0.01 and 0.67 ± 0.01, respectively). None of the combined features outperformed the SVM using clinical features.

 doi:10.20944/preprints202212.0195.v1

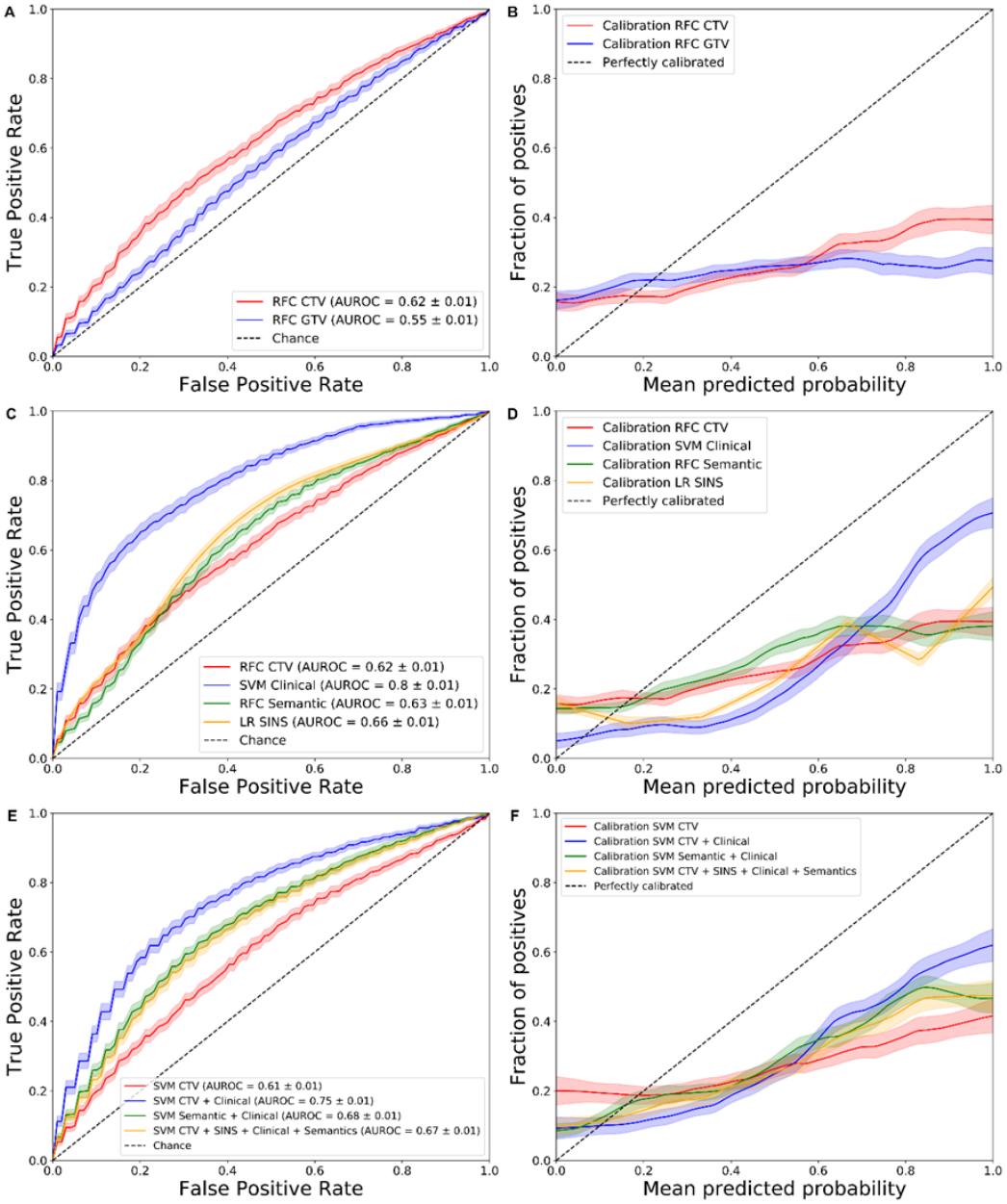


Figure 2. Receiver operator characteristic (ROC) and Calibration curves for the comparisons of different segmentation modes (A, B), clinical baseline, semantic and SINS features (C, D), and combined models (E, F).

3.5. Feature importance

Feature importance was estimated for SVM and RFC trained on CTV, clinical baseline, semantic, and combined sets of data (Tables S8 to S11). None of the features from the CTV models were selected in all of the 250 cases. On the other hand, the top 15 features for both CTV models (SVM and RFC) were highly homogeneous, sharing the same top three texture features. The most important semantic features were the extent of the GTV along with features also used in the SINS score (e.g., lytic bone lesions and bilateral posterolateral involvement of the spinal element).

The mean decrease in impurity of the RFC nodes, overall, showed low values, with most being below 0.1. However, clinical features achieved a significantly higher feature importance, which is in concordance with the higher performance of those models.

The combined SVM model of CTV, Clinical, SINS and Semantic features showed the same low importance values, with almost no feature selected in 100% of the cases. The

 doi:10.20944/preprints202212.0195.v1

feature that was selected most often, while also retaining high importance, was the clinical feature "Tumour Type: Breast Cancer" followed by predominantly semantic and clinical features. Although the majority of all features in the combined model were radiomic (105 of 135), only four of the 12 most predictive features were radiomic, while most of them were semantic.

4. Discussion

In this exploratory analysis we analysed the potential of ML models to predict pain response to RT of PSBM. CT-based radiomic machine learning models predicted pain response better than random. CTV-based outperformed GTV-based models; semantic and SINS-based models outperformed random, and clinical models performed best, with SVM at the peak. The combination of radiomic features with clinical data significantly increased performance compared to the radiomic baseline. This combination, however, did not match models using only clinical features. The addition of the SINS feature neither affected the radiomics nor the combined model. The feature importance of all radiomic features showed low levels of mean impurity decrease in RFC. Texture features were selected as most important predictors. Only clinical features have shown a high importance level, while they were also often consistently selected.

In our modelling approach, we compared two established ML models. Both models achieved competitive results. The best ML model, for radiomics data, was RFC by a small but statistically significant margin (Tables 3 and S4). However, the SVM performed better in some situations, mainly for other metrics such as the BA and F1 score. In addition, the SVM achieved the best results when trained on clinical data, and performed better than the RFC for the combined data models (Table 5 and S6). The SVM achieved more stable results when trained on radiomics features, with more competent prediction qualities when trained on features with, in principle, less useful information. Given the low importance of these features, these results indicate that the SVM is more resilient to selected features with poor importance. This is further confirmed when analysing the combined models: combined SVM models achieved consistently better performances than RFCs.

We have compared the predictive performance of multiple sets of data: two radiomic segmentation modes, clinical, semantic and SINS features. The only model that did not achieve better than random results was LR trained only on SINS (binarized; Table 4). This is to be expected: by binarizing the SINS variable, important information, that can be learnt by either model, is lost. Combined models that used clinical data had an expected performance increase compared to their respective baselines (Table 5 and S6). However, these combined models performed worse than a clinical only model: this indicates that the addition of features that are not important to the model can have a negative impact on its performance, by making it difficult for the model to identify patterns in the data. This is further confirmed by the decrease in feature importance of the clinical features when comparing them alone and in a combined model (Tables S9 and S11, respectively).

All radiomic features have shown low feature importance, which can be explained by a possible low correlation to the prediction target. This is also consistent with the fact that none were selected in any of the 250 cases. In addition, only 10 of all 105 features were selected at least 50% of the time, which indicates high variance when selecting features, possibly due to their low correlation. This is not the case for clinical features, which have shown more than thrice higher feature importance, and were selected in nearly all cases when used in combined modelling (Tables S9 and S11).

Multiple previous publications have analysed factors related to pain response following RT of bone metastasis. An early retrospective study by Arcangeli et al. demonstrated that pain response depended on patients' performance status and specific histology. NSCLC patients were shown to have a worse response to RT than patients with other cancer origins [32]. This was reproduced by Nyguen et al. demonstrating a favourable response for patients with prostate and mammary carcinoma [33]. Location and pain level before therapy appeared not to influence radiation response [40,41]. These results were

 doi:10.20944/preprints202212.0195.v1

validated in a large prospective trial with 956 patients by Westhoff et al. Next to the aforementioned clinical factors, the use of opioids and absence of visceral metastases were positively predictive for RT response [30]. However, the multivariate model achieved only limited predictive capacity with a C-statistic of 0.56. Van Velden et al. conducted a further prospective trial comparing the predictive performance of the SINS with clinical parameters [29]. SINS appeared to be significantly associated with complete response after adjustment for gender, tumour type and performance status. Adding SINS to the clinical parameters increased the AUROC for the prediction of complete response from 0.68 to 0.78. In our study, SINS as training data proved to perform better than random (Table 4). However, adding SINS to other datasets did not increase their performance significantly (Table 5 and S6). Combining clinical and SINS data, the overall performance was significantly better than the radiomic models (SVM and RFC AUROCS: 0.73 ± 0.01 and 0.75 ± 0.01, respectively), albeit inferior to the clinical models (Table S7).

In our study, we compared two potential modes of segmentation. Although the predictive performance was overall similar, the CTV-based segmentations were superior for both ML models. In contrast to the GTV, the CTV segmentation included vertebra compartments that are at risk of microscopic infiltration [35]. This additional information may have improved the predictive power. Texture features were the most important radiomic features. Such features may capture texture and intensity heterogeneity that may be associated with cell density within the bone marrow. Analysis of magnetic resonance imaging data may be more suitable to quantify such changes. Recently, one other publication has analysed the potential of radiomics-based prediction of pain response [42]. The authors trained a random forest model on a single centre cohort of 69 patients using leave-one-out cross-validation. While their clinical model showed an inferior performance with an AUC of 0.70, the radiomic model was able to predict pain response with a superior AUC of 0.82. There are several reasons that may explain these differences in performance. First, the authors applied only the simplistic double-layer split into train and test set (through their leave-one-out cross-validation), instead of the more adequate triple-layer split into train, validation and test set. Consequently, the authors optimized their model on the same patients used for assessing performance, thereby opening the door to data leakage. Such leakage often leads to substantial over-estimates of performance. In contrast, our nested cross-validation results included repeated testing independent of hyperparameter optimization guaranteeing more unbiased results. Second, the authors used a different set of VOIs. Instead of a GTV or specific CTV, the authors used the spinal canal, the complete vertebra and the vertebra plus a one-centimetre margin as VOIs. So far, it remains unclear what segmentation strategy may be optimal. Third, the authors trained their model for "any pain response" instead of "complete response" which may also explain a difference in performance by having a broader prediction target. Taken together with our results, both studies could demonstrate prediction of pain response better than random albeit with different predictive power against the background of a significantly different study design.

Besides quantitative radiomic image analysis, semantic feature extraction constitutes an alternative "manual" way to extract information from medical images [43]. For prediction of pain response of PSBM Mitera et al. evaluated semantic imaging features in 33 patients [44]. The authors did not find any association of semantic imaging features to pain response. Semantic features included pathological fractures, kyphosis and anatomic extent of tumour. However, the study was limited by the use of a large number of semantic features and a relatively small number of patients. For instance, the known predictive factor age did not correlate with response either. Our study has shown that, with a larger training set, it is possible to achieve better than random prediction results when training either ML algorithm with semantic data, with RFC performing best (AUROC: 0.63 ± 0.01; Table 4). It is important to note that the SINS score in itself is a score combining multiple semantic features. We used these features complemented with other additional variables.

 doi:10.20944/preprints202212.0195.v1

The SINS score, however, performed better than the semantic model, demonstrating that the important features are already included in the SINS score.

There are several limitations to our study. First, pain response was assessed retrospectively. Due to non-standardized or incomplete reporting of pain response determination, it may have been error prone. To allow a standardised assessment we followed the recommendation of the International Spine Radiosurgery Consortium Consensus Guidelines [35]. Patients with "indeterminate response" were excluded from analysis. This may have conferred a selection bias as missing information may be associated with confounding factors such as low KPS or early death. Secondly, in patients with multiple PBMS each metastasis was treated as a separate sample. The outcome, however, was equal between all metastases of a specific patient. Information on which specific metastases contributed to symptomatic pain remained elusive. To prevent data leakage and bias, stratified cross-validation was performed, guaranteeing that multiple samples from the same patient were evenly distributed across all splits. Thirdly, our study was of monocentric nature with a lack of an external validation set. To compensate for this, we applied nested cross-validation and repeated the process 50 times to increase the statistical strength of the results. We believe that our exploratory analysis allows the assessment of the general possibility of RT response prediction and a comparison to established factors.

5. Conclusions

To conclude, in this exploratory work we were able to demonstrate a strong predictive value of machine learning based prediction of complete pain response to palliative radiotherapy in patients with painful spinal bone metastases using established clinical factors. CT-based radiomics and semantic machine learning models performed better than random but sub-optimally. The SINS score performed slightly better than both, and models trained on a combination of the available datasets performed even better. Using exclusively clinical features as input, however, outperformed all other models. Upon inspection of the radiomic and clinical features, their importance and selection frequency confirmed the higher predictive quality of the latter, with a more than three-fold decrease in mean impurity. As a consequence, appeared to be no need for CT-based image analysis to predict pain response.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1; Table S1: Histology distribution, Table S2: CT acquisition Parameters, Table S3: Extracted radiomics features, Table S4: Extension of Table 3, Table S5: Extension of Table 4, Table S6: Extension of Table 5, Table S7: AUROC, BA, F1 Score and MCC for the SVM and RFC models trained on clinical and SINS features, Table S8: Feature importance table for SVM and RFC models trained on CTV features, Table S9: Feature importance table for SVM and RFC models trained on clinical features, Table S10: Feature importance table for SVM and RFC models trained on semantic features, Table S11: Feature importance table for SVM and RFC models trained on CTV, clinical, SINS and semantic features, Figure S1: Patients workflow, Figure S2: ROC and Calibration Curves: Extension of Figure 2.

Author Contributions: Conceptualization, Steffi Pigorsch and Jan Peeken; Data curation, Joachim Akhgar and Jan Peeken; Formal analysis, Oscar Llorián-Salvador, Joachim Akhgar and Jan Peeken; Funding acquisition, Stephanie Combs and Jan Peeken; Investigation, Oscar Llorián-Salvador and Jan Peeken; Methodology, Oscar Llorián-Salvador and Jan Peeken; Project administration, Jan Peeken; Resources, Denise Bernhardt, Burkhard Rost and Stephanie Combs; Software, Oscar Llorián-Salvador; Supervision, Burkhard Rost, Miguel Andrade-Navarro, Stephanie Combs and Jan Peeken; Validation, Oscar Llorián-Salvador and Jan Peeken; Visualization, Oscar Llorián-Salvador and Joachim Akhgar; Writing – original draft, Oscar Llorián-Salvador and Jan Peeken; Writing – review & editing, Joachim Akhgar, Steffi Pigorsch, Kai Borm, Stefan Münch, Denise Bernhardt, Burkhard Rost, Miguel Andrade-Navarro and Stephanie Combs. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by physician scientist programs of the medical faculty of the Technical University of Munich and the Helmholtz Zentrum Muenchen. Funding was also received from Else-Kröner-Fresenius-Stiftung

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Simes, R.J. Treatment Selection for Cancer Patients: Application of Statistical Decision Theory to the Treatment of Advanced Ovarian Cancer. *J Chronic Dis* **1985**, *38*, 171–186.
2. Maclin, P.S.; Dempsey, J.; Brooks, J.; Rand, J. Using Neural Networks to Diagnose Cancer. *J Med Syst* **1991**, *15*, 11–19.
3. Cicchetti, D.V. Neural Networks and Diagnosis in the Clinical Laboratory: State of the Art. *Clin Chem* **1992**, *38*, 9–10.
4. Koza, J.R.; Bennett, F.H.; Andre, D.; Keane, M.A. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In *Artificial Intelligence in Design '96*; Gero, J.S., Sudweeks, F., Eds.; Springer Netherlands: Dordrecht, 1996; pp. 151–170 ISBN 978-94-009-0279-4.
5. Mitchell, T.M. *Machine Learning*; 1st ed.; McGraw-Hill, Inc.: USA, 1997; ISBN 978-0-07-042807-2.
6. Peeken, J.C.; Goldberg, T.; Knie, C.; Komboz, B.; Bernhofer, M.; Pasa, F.; Kessel, K.A.; Tafti, P.D.; Rost, B.; Nüsslin, F.; et al. Treatment-Related Features Improve Machine Learning Prediction of Prognosis in Soft Tissue Sarcoma Patients. *Strahlenther Onkol* **2018**, *194*, 824–834, doi:10.1007/s00066-018-1294-2.
7. Peeken, J.C.; Goldberg, T.; Pyka, T.; Bernhofer, M.; Wiestler, B.; Kessel, K.A.; Tafti, P.D.; Nüsslin, F.; Braun, A.E.; Zimmer, C.; et al. Combining Multimodal Imaging and Treatment Features Improves Machine Learning-Based Prognostic Assessment in Patients with Glioblastoma Multiforme. *Cancer Medicine* **2019**, *8*, 128–136, doi:10.1002/cam4.1908.
8. Peeken, J.C.; Spraker, M.B.; Knebel, C.; Dapper, H.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; Eisenhart-Rothe, R. von; et al. Tumor Grading of Soft Tissue Sarcomas Using MRI-Based Radiomics. *eBioMedicine* **2019**, *48*, 332–340, doi:10.1016/j.ebiom.2019.08.059.
9. Gupta, S.; Tran, T.; Luo, W.; Phung, D.; Kennedy, R.L.; Broad, A.; Campbell, D.; Kipp, D.; Singh, M.; Khasraw, M.; et al. Machine-Learning Prediction of Cancer Survival: A Retrospective Study Using Electronic Administrative Records and a Cancer Registry. *BMJ Open* **2014**, *4*, e004007, doi:10.1136/bmjopen-2013-004007.
10. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The Bridge between Medical Imaging and Personalized Medicine. *Nat Rev Clin Oncol* **2017**, *14*, 749–762, doi:10.1038/nrclinonc.2017.141.
11. Peeken, J.C.; Bernhofer, M.; Wiestler, B.; Goldberg, T.; Cremers, D.; Rost, B.; Wilkens, J.J.; Combs, S.E.; Nüsslin, F. Radiomics in Radiooncology - Challenging the Medical Physicist. *Physica Medica: European Journal of Medical Physics* **2018**, *48*, 27–36, doi:10.1016/j.ejmp.2018.03.012.
12. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur J Cancer* **2012**, *48*, 441–446, doi:10.1016/j.ejca.2011.11.036.
13. Peeken, J.C.; Nusslin, F.; Combs, S.E. "Radio-Oncomics" : The Potential of Radiomics in Radiation Oncology. *Strahlenther Onkol* **2017**, *193*, 767–779, doi:10.1007/s00066-017-1175-0.
14. Peeken, J.C.; Wiestler, B.; Combs, S.E. Image-Guided Radiooncology: The Potential of Radiomics in Clinical Application. *Recent Results Cancer Res* **2020**, *216*, 773–794, doi:10.1007/978-3-030-42618-7_24.
15. Peeken, J.C.; Neumann, J.; Asadpour, R.; Leonhardt, Y.; Moreira, J.R.; Hippe, D.S.; Klymenko, O.; Foreman, S.C.; von Schacky, C.E.; Spraker, M.B.; et al. Prognostic Assessment in High-Grade Soft-Tissue Sarcoma Patients: A Comparison of Semantic Image Analysis and Radiomics. *Cancers* **2021**, *13*, 1929, doi:10.3390/cancers13081929.
16. Peeken, J.C.; Bernhofer, M.; Spraker, M.B.; Pfeiffer, D.; Devecka, M.; Thamer, A.; Shouman, M.A.; Ott, A.; Nüsslin, F.; Mayr, N.A.; et al. CT-Based Radiomic Features Predict Tumor Grading and Have Prognostic Value in Patients with Soft Tissue Sarcomas Treated with Neoadjuvant Radiation Therapy. *Radiotherapy and Oncology* **2019**, *135*, 187–196, doi:10.1016/j.radonc.2019.01.004.
17. Peeken, J.C.; Asadpour, R.; Specht, K.; Chen, E.Y.; Klymenko, O.; Akinkuoroye, V.; Hippe, D.S.; Spraker, M.B.; Schaub, S.K.; Dapper, H.; et al. MRI-Based Delta-Radiomics Predicts Pathologic Complete Response in High-Grade Soft-Tissue Sarcoma Patients Treated with Neoadjuvant Therapy. *Radiotherapy and Oncology* **2021**, *164*, 73–82, doi:10.1016/j.radonc.2021.08.023.
18. Peeken, J.C.; Shouman, M.A.; Kroenke, M.; Rauscher, I.; Maurer, T.; Gschwend, J.E.; Eiber, M.; Combs, S.E. A CT-Based Radiomics Model to Detect Prostate Cancer Lymph Node Metastases in PSMA Radioguided Surgery Patients. *Eur J Nucl Med Mol Imaging* **2020**, *47*, 2968–2977, doi:10.1007/s00259-020-04864-1.
19. Peeken, J.C.; Molina-Romero, M.; Diehl, C.; Menze, B.H.; Straube, C.; Meyer, B.; Zimmer, C.; Wiestler, B.; Combs, S.E. Deep Learning Derived Tumor Infiltration Maps for Personalized Target Definition in Glioblastoma Radiotherapy. *Radiotherapy and Oncology* **2019**, *138*, 166–172, doi:10.1016/j.radonc.2019.06.031.
20. Lang, D.M.; Peeken, J.C.; Combs, S.E.; Wilkens, J.J.; Bartzsch, S. Deep Learning Based HPV Status Prediction for Oropharyngeal Cancer Patients. *Cancers* **2021**, *13*, 786, doi:10.3390/cancers13040786.

21. Navarro, F.; Dapper, H.; Asadpour, R.; Knebel, C.; Spraker, M.B.; Schwarze, V.; Schaub, S.K.; Mayr, N.A.; Specht, K.; Woodruff, H.C.; et al. Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. *Cancers* **2021**, *13*, 2866, doi:10.3390/cancers13122866.
22. Leger, S.; Zwanenburg, A.; Leger, K.; Lohaus, F.; Linge, A.; Schreiber, A.; Kalinauskaite, G.; Tinhofer, I.; Guberina, N.; Guberina, M.; et al. Comprehensive Analysis of Tumour Sub-Volumes for Radiomic Risk Modelling in Locally Advanced HNSCC. *Cancers* **2020**, *12*, 3047, doi:10.3390/cancers12103047.
23. Starke, S.; Leger, S.; Zwanenburg, A.; Leger, K.; Lohaus, F.; Linge, A.; Schreiber, A.; Kalinauskaite, G.; Tinhofer, I.; Guberina, N.; et al. 2D and 3D Convolutional Neural Networks for Outcome Modelling of Locally Advanced Head and Neck Squamous Cell Carcinoma. *Sci Rep* **2020**, *10*, 15625, doi:10.1038/s41598-020-70542-9.
24. Marr, L.; Haller, B.; Pyka, T.; Peeken, J.C.; Jesinghaus, M.; Scheidhauer, K.; Friess, H.; Combs, S.E.; Münch, S. Predictive Value of Clinical and 18F-FDG-PET/CT Derived Imaging Parameters in Patients Undergoing Neoadjuvant Chemoradiation for Esophageal Squamous Cell Carcinoma. *Sci Rep* **2022**, *12*, 7148, doi:10.1038/s41598-022-11076-0.
25. Spohn, S.K.B.; Farolfi, A.; Schandeler, S.; Vogel, M.M.E.; Ruf, J.; Mix, M.; Kirste, S.; Ceci, F.; Fanti, S.; Lanzafame, H.; et al. The Maximum Standardized Uptake Value in Patients with Recurrent or Persistent Prostate Cancer after Radical Prostatectomy and PSMA-PET-Guided Salvage Radiotherapy—a Multicenter Retrospective Analysis. *Eur J Nucl Med Mol Imaging* **2022**, doi:10.1007/s00259-022-05931-5.
26. Shahzadi, I.; Zwanenburg, A.; Lattermann, A.; Linge, A.; Baldus, C.; Peeken, J.C.; Combs, S.E.; Diefenhardt, M.; Rödel, C.; Kirste, S.; et al. Analysis of MRI and CT-Based Radiomics Features for Personalized Treatment in Locally Advanced Rectal Cancer and External Validation of Published Radiomics Models. *Sci Rep* **2022**, *12*, 10192, doi:10.1038/s41598-022-13967-8.
27. Dijk, L.V. van; Thor, M.; Steenbakkers, R.J.H.M.; Apte, A.; Zhai, T.-T.; Borra, R.; Noordzij, W.; Estilo, C.; Lee, N.; Langendijk, J.A.; et al. Parotid Gland Fat Related Magnetic Resonance Image Biomarkers Improve Prediction of Late Radiation-Induced Xerostomia. *Radiotherapy and Oncology* **2018**, *128*, 459–466, doi:10.1016/j.radonc.2018.06.012.
28. Krafft, S.P.; Rao, A.; Stingo, F.; Briere, T.M.; Court, L.E.; Liao, Z.; Martel, M.K. The Utility of Quantitative CT Radiomics Features for Improved Prediction of Radiation Pneumonitis. *Medical Physics* **2018**, *45*, 5317–5324, doi:10.1002/mp.13150.
29. van der Velden, J.M.; Versteeg, A.L.; Verkooijen, H.M.; Fisher, C.G.; Chow, E.; Oner, F.C.; van Vulpen, M.; Weir, L.; Verlaan, J. Prospective Evaluation of the Relationship Between Mechanical Stability and Response to Palliative Radiotherapy for Symptomatic Spinal Metastases. *Oncologist* **2017**, *22*, 972–978, doi:10.1634/theoncologist.2016-0356.
30. Westhoff, P.G.; Graeff, A. de; Monninkhof, E.M.; Pomp, J.; Vulpen, M. van; Leer, J.W.H.; Marijnen, C.A.M.; Linden, Y.M. van der Quality of Life in Relation to Pain Response to Radiation Therapy for Painful Bone Metastases. *International Journal of Radiation Oncology, Biology, Physics* **2015**, *93*, 694–701, doi:10.1016/j.ijrobp.2015.06.024.
31. A Novel Classification System for Spinal Instability in Neoplastic Disease: An Evidence-Based Approach and Expert Consensus From the Spine Oncology Study Group Available online: https://oce.ovid.com/article/00007632-201010150-00019/HTML (accessed on 21 October 2022).
32. Arcangeli, G.; Giovinazzo, G.; Saracino, B.; D'Angelo, L.; Giannarelli, D.; Arcangeli, G.; Micheli, A. Radiation Therapy in the Management of Symptomatic Bone Metastases: The Effect of Total Dose and Histology on Pain Relief and Response Duration. *Int J Radiat Oncol Biol Phys* **1998**, *42*, 1119–1126, doi:10.1016/s0360-3016(98)00264-8.
33. Nguyen, J.; Chow, E.; Zeng, L.; Zhang, L.; Culleton, S.; Holden, L.; Mitera, G.; Tsao, M.; Barnes, E.; Danjoux, C.; et al. Palliative Response and Functional Interference Outcomes Using the Brief Pain Inventory for Spinal Bony Metastases Treated with Conventional Radiotherapy. *Clinical Oncology* **2011**, *23*, 485–491, doi:10.1016/j.clon.2011.01.507.
34. Chow, E.; Hoskin, P.; Mitera, G.; Zeng, L.; Lutz, S.; Roos, D.; Hahn, C.; Linden, Y. van der; Hartsell, W.; Kumar, E. Update of the International Consensus on Palliative Radiotherapy Endpoints for Future Clinical Trials in Bone Metastases. *International Journal of Radiation Oncology, Biology, Physics* **2012**, *82*, 1730–1737, doi:10.1016/j.ijrobp.2011.02.008.
35. Cox, B.W.; Spratt, D.E.; Lovelock, M.; Bilsky, M.H.; Lis, E.; Ryu, S.; Sheehan, J.; Gerszten, P.C.; Chang, E.; Gibbs, I.; et al. International Spine Radiosurgery Consortium Consensus Guidelines for Target Volume Definition in Spinal Stereotactic Radiosurgery. *International Journal of Radiation Oncology, Biology, Physics* **2012**, *83*, e597–e605, doi:10.1016/j.ijrobp.2012.03.009.
36. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **2017**, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
37. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338, doi:10.1148/radiol.2020191145.
38. Ding, C.; Peng, H. Minimum Redundancy Feature Selection From Microarray Gene Expression Data.; September 11 2003; Vol. 3, pp. 523–528.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
40. Kirou-Mauro, A.; Hird, A.; Wong, J.; Sinclair, E.; Barnes, E.A.; Tsao, M.; Danjoux, C.; Chow, E. Is Response to Radiotherapy in Patients Related to the Severity of Pretreatment Pain? *International Journal of Radiation Oncology, Biology, Physics* **2008**, *71*, 1208–1212, doi:10.1016/j.ijrobp.2007.11.062.

41. Zeng, L.; Chow, E.; Zhang, L.; Culleton, S.; Holden, L.; Jon, F.; Khan, L.; Tsao, M.; Barnes, E.; Danjoux, C.; et al. Comparison of Pain Response and Functional Interference Outcomes between Spinal and Non-Spinal Bone Metastases Treated with Palliative Radiotherapy. *Support Care Cancer* **2012**, *20*, 633–639, doi:10.1007/s00520-011-1144-6.
42. Wakabayashi, K.; Koide, Y.; Aoyama, T.; Shimizu, H.; Miyauchi, R.; Tanaka, H.; Tachibana, H.; Nakamura, K.; Kodaira, T. A Predictive Model for Pain Response Following Radiotherapy for Treatment of Spinal Metastases. *Sci Rep* **2021**, *11*, 12908, doi:10.1038/s41598-021-92363-0.
43. Peeken, J.C.; Hesse, J.; Haller, B.; Kessel, K.A.; Nüsslin, F.; Combs, S.E. Semantic Imaging Features Predict Disease Progression and Survival in Glioblastoma Multiforme Patients. *Strahlenther Onkol* **2018**, *194*, 580–590, doi:10.1007/s00066-018-1276-4.
44. Mitera, G.; Probyn, L.; Ford, M.; Donovan, A.; Rubenstein, J.; Finkelstein, J.; Christakis, M.; Zhang, L.; Campos, S.; Culleton, S.; et al. Correlation of Computed Tomography Imaging Features with Pain Response in Patients with Spine Metastases after Radiation Therapy. *Int J Radiat Oncol Biol Phys* **2011**, *81*, 827–830, doi:10.1016/j.ijrobp.2010.06.036.