frontiers in Oncology

ORIGINAL RESEARCH
published: 03 December 2015
doi: 10.3389/fonc.2015.00272

Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer

Chintan Parmar[1,2,3]†, Patrick Grossmann[1,2,4]†, Derek Rietveld[5], Michelle M. Rietbergen[6], Philippe Lambin[3] and Hugo J. W. L. Aerts[1,2,4]**

[1] Department of Radiation Oncology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [2] Department of Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [3] Radiation Oncology (MAASTRO), Research Institute GROW, Maastricht University, Maastricht, Netherlands, [4] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA, [5] Department of Radiation Oncology, VU University Medical Center, Amsterdam, Netherlands, [6] Department of Otolaryngology/Head and Neck Surgery, VU University Medical Center, Amsterdam, Netherlands

Introduction: "Radiomics" extracts and mines a large number of medical imaging features in a non-invasive and cost-effective way. The underlying assumption of radiomics is that these imaging features quantify phenotypic characteristics of an entire tumor. In order to enhance applicability of radiomics in clinical oncology, highly accurate and reliable machine-learning approaches are required. In this radiomic study, 13 feature selection methods and 11 machine-learning classification methods were evaluated in terms of their performance and stability for predicting overall survival in head and neck cancer patients.

Methods: Two independent head and neck cancer cohorts were investigated. Training cohort HN1 consisted of 101 head and neck cancer patients. Cohort HN2 ($n = 95$) was used for validation. A total of 440 radiomic features were extracted from the segmented tumor regions in CT images. Feature selection and classification methods were compared using an unbiased evaluation framework.

Results: We observed that the three feature selection methods minimum redundancy maximum relevance (AUC = 0.69, Stability = 0.66), mutual information feature selection (AUC = 0.66, Stability = 0.69), and conditional infomax feature extraction (AUC = 0.68, Stability = 0.7) had high prognostic performance and stability. The three classifiers BY (AUC = 0.67, RSD = 11.28), RF (AUC = 0.61, RSD = 7.36), and NN (AUC = 0.62, RSD = 10.52) also showed high prognostic performance and stability. Analysis investigating performance variability indicated that the choice of classification method is the major factor driving the performance variation (29.02% of total variance).

Conclusion: Our study identified prognostic and reliable machine-learning methods for the prediction of overall survival of head and neck cancer patients. Identification of optimal machine-learning methods for radiomics-based prognostic analyses could broaden the scope of radiomics in precision oncology and cancer care.

Keywords: quantitative imaging, radiology, radiomics, cancer, machine learning, computational science

OPEN ACCESS

Edited by:
Issam El Naqa,
McGill University, Canada

Reviewed by:
Kevin Du,
NYU Langone Medical Center, USA
X. Allen Li,
Medical College of Wisconsin, USA

****Correspondence:***
Chintan Parmar
chintan_parmar@dfci.harvard.edu;
Hugo J. W. L. Aerts
hugo_aerts@dfci.harvard.edu

†Chintan Parmar and Patrick Grossmann have contributed equally to this work.

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *09 September 2015*
Accepted: *20 November 2015*
Published: *03 December 2015*

Citation:
Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P and Aerts HJWL (2015) Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Front. Oncol. 5:272. doi: 10.3389/fonc.2015.00272

INTRODUCTION

The emergence of "radiomics" (1) has expanded the scope of medical imaging in clinical oncology. Radiomics focuses on extracting and mining a large number of medical imaging features. It is hypothesized that these imaging features are enriched with crucial information regarding tumor phenotype (1, 2). These features provide a comprehensive characterization of entire tumors, and hence are likely to capture the intra-tumor heterogeneity. It has been stated that intra-tumor heterogeneity could have profound implications in clinical predictions (e.g., treatment response, survival outcomes, disease progression, etc.), and therefore it is considered as a crucial factor for precision oncology and related research (3–6). Several studies have assessed various radiomic features in different cancer types and with respect to different imaging modalities (2, 7–11). Some studies have investigated the reproducibility/variability of radiomic features across different clinical settings (2, 12–14). Moreover, several other studies have reported significant predictive/prognostic power of radiomic features. It has been shown that radiomic features are associated with tumor histology (15–17), tumor grades or stages (16), patient survival (2, 7, 18–20), metabolism (21), and various other clinical outcomes (7, 16, 22, 23). Furthermore, some radio-genomic studies have reported associations between radiomic features and underlying gene expression patterns (2, 9, 11, 24, 25). These reports indicate that radiomics could improve individualized treatment selection and monitoring. Furthermore, unlike most of the genomic-based approaches, radiomics is non-invasive and relatively cost-effective (2, 26). Therefore, radiomics is a novel and promising step forward toward the realization of precision oncology.

Predictive and prognostic models are an important part of radiomics (27). Highly accurate and reliable models are desired to improve decision support in clinical oncology. Machine learning could help in this regard. Machine learning can be broadly defined as computational methods/models using data to improve performance or make accurate predictions (28). These programmable methods can "learn" from the data, and hence automate and improve the prediction process. Therefore, it is essential to compare different machine-learning models for precision oncology, and hence also for radiomics-based clinical biomarkers. Recent advances in medical image acquisition technologies allow higher resolution tumor imaging and facilitate detailed quantification of tumor phenotype. The feature dimensions of radiomics are increasing rapidly. One of the issues with high dimensional feature space is the "curse of dimensionality" (29). A large number of features with limited sample size could hinder the predictive/prognostic power of a model. Feature/variable selection is one of the ways to tackle the curse of dimensionality. Therefore, different feature selection methods (29) should be thoroughly investigated for radiomics-based prognostic analyses. However, as radiomics is an emerging field of research, not sufficient effort could be made toward assessing the impact of different machine-learning methods. The majority of the radiomics-based studies have only assessed the discriminating power of radiomic features without evaluating alternative prediction/prognostic models.

Only few recent studies have compared different feature selection and classification methods on radiomics-based clinical predictions (15, 20), but with limited sample sizes and also without independent validation. In a recently published radiomic study (30), a large panel of feature selection and machine-learning classification methods was evaluated in two independent cohorts of patients with non-small cell lung cancer (NSCLC). They proposed an unbiased framework for comparing different feature selection and classification methods using publicly available implementations (31, 32) and reported parameter configurations (33).

In this study, we assessed a large panel of machine-learning methods for overall survival prediction of head and neck cancer (HNSCC) patients. Two independent HNSCC cohorts totaling 196 patients were used in the analysis. Feature selection and classification training was done using training cohort HN1 and the prediction performance was evaluated in the validation cohort HN2. All the feature selection and classification methods were evaluated in terms of their prognostic ability and stability against data perturbation. Machine-learning methods having high prognostic/predictive power and stability are desired for radiomics-based analyses. Such methods could enhance the applications of non-invasive and cost-effective radiomics in cancer care.

MATERIALS AND METHODS

Radiomic Features

We used 440 radiomic features describing the first order intensity statistics, texture (34, 35), and shape of the three-dimensional tumor region on CT images. Intensity and textural features were also recomputed after different wavelet decomposition of the original image. Mathematical definitions of all radiomic features as well as the extraction methods were previously described (2, 30).

Datasets

In this study, we analyzed two HNSCC cohorts from the two different institutes of Netherlands:

(1) HN1: 136 HNSCC patients treated at MAASTRO Clinic, Maastricht. All patients received a treatment planning ^{18F}FDG-PET-CT scan (Biograph, SOMATOM Sensation-16 with an ECAT ACCEL PET scanner; Siemens, Erlangen, Germany) made with the patient immobilized using a thermoplastic mask. Patients fasted at least 6 hours before the start of the acquisition. A total dose dependent on the weight of the patient (weight ×4 + 20 MBq) of [18F] fluoro-2-deoxy-D-glucose (FDG) 30 (MDS Nordion, Liège, Belgium), was injected intravenously, followed by physiological saline (10 mL). Free-breathing PET and CT images were acquired after an uptake period of 45 min. A spiral CT (3 mm slice thickness) was performed covering the complete thoracic region. Based on the radiological examinations and clinical findings, the gross tumor volume (GTV) was delineated on the fused PET-CT scan by a radiation oncologist in a radiotherapy treatment planning system (XiO, CMS, St Louis, MO, USA).

(2) HN2: 95 HNSCC patients treated at VU University Medical Center (VUMC), Amsterdam. All patients received a treatment planning CT scan of the head and neck (Varian Medical Systems VISION 3253). CT scans were acquired in helical mode with slice thickness of 2.5 mm. The GTV was delineated by an experienced radiation oncologist on the CT scans.

This analysis was carried out in accordance with Dutch law. The Institutional Review Boards of both the participating centers approved the studies: HN1 (MAASTRO Clinic Maastricht, The Netherlands) and HN2 (VUMC, Amsterdam, The Netherlands). Further details related to patient population and treatments can be obtained from the previous study (2). We dichotomized the censored continuous survival data using a cutoff time of 3 years. Patients who survived beyond the cutoff time were labeled as 1, whereas the deceased ones were labeled as 0. The objective of the study was to stratify patients into these two labeled survival classes using different machine-learning classifiers. We used 3 years survival cut-off because it resulted in reasonable event ratios (37% for HN1, 34% for HN2) in the cohorts. We excluded patients that were followed up for <3 years. This resulted in 101 patients in the training cohort (HN1) and 95 patients in the validation cohort (HN2). All features were standardized using *Z*-score standardization.

Feature Selection Methods

As described in a previously published study (30), 14 FS methods based on filter approaches were used in the analysis: Fisher score (FSCR), relief (RELF), *T*-score (TSCR), Chi-square (CHSQ), Wilcoxon (WLCX), Gini index (GINI), mutual information maximization (MIM), mutual information feature selection (MIFS), minimum redundancy maximum relevance (MRMR), conditional infomax feature extraction (CIFE), joint mutual information (JMI), conditional mutual information maximization (CMIM), interaction capping (ICAP), and double input symmetric relevance (DISR). However, the method CHSQ did not run according to our experimental design. CHSQ was not able to select the required number of features due to the smaller size of training cohort. We thus removed it from further analysis. The acronyms related to the feature selection methods are defined in **Table 1**. Publicly available implementations were used for these methods (31, 32). Detailed description regarding these methods can be obtained from Parmar et al. (30).

Classifiers

As described earlier (30), we investigated 12 machine-learning classifiers belonging to the 12 classifier families: bagging (BAG), Bayesian (BY), boosting (BST), decision trees (DT), discriminant analysis (DA), generalized linear models (GLM), multiple adaptive regression splines (MARS), nearest neighbors (NN), neural networks (Nnet), partial least square and principle component regression (PLSR), random forests (RF), and support vector machines (SVM). In our experimental settings, classifier DA generated computation error in the majority of cases. This could be due to the smaller training cohort. Therefore, we removed DA from further analysis and used the remaining 11 classifiers. The acronyms related to the classifiers are defined in **Table 1**. All classifiers were implemented using the R package caret (version 6.0-47) (36), which provides a nice interface to access many machine-learning algorithms in R. Classifiers were trained using the repeated (three repeat iterations) 10-fold cross validation in the training cohort (HN1), and their prognostic performance was evaluated in the validation cohort (HN2) using the area under receiver operator characteristic (ROC) curve (AUC). We used the classifier parameters as defined by earlier studies (30, 33). All the classifiers, the corresponding parameters and R packages are listed in Ref. (30).

TABLE 1 | Table defining the acronyms related to the used feature selection and classification methods.

Classification method acronym	Classification method name	Feature Selection method acronym	Feature selection method name
Nnet	Neural network	RELF	Relief
DT	Decision tree	FSCR	Fisher score
BST	Boosting	GINI	Gini index
BY	Bayesian	JMI	Joint mutual information
BAG	Bagging	CIFE	Conditional infomax feature extraction
RF	Random forset	DISR	Double input symmetric relevance
MARS	Multi adaptive regression splines	MIM	Mutual information maximization
SVM	Support vector machines	CMIM	Conditional mutual information maximization
NN	Neirest neighbor	ICAP	Interaction capping
GLM	Generalized linear models	TSCR	*T*-test score
PLSR	Partial least squares and prinicipal component regression	MRMR	Minimum redundancy maximum relevance
–	–	MIFS	Mutual information feature selection
–	–	WLCX	Wilcoxon

Analysis

Prognostic Performance

We compared different feature selection and classification methods using the experimental design defined by an earlier study of NSCLC radiomic cohorts (30). We incrementally selected features ranging from 5 up to 50, with an increment of 5 features ($n = 5, 10, 15, 20, \dots, 50$), using each of the 13 feature selection methods. These subsets of selected features were then used as an input to each of the 11 machine-learning classifiers. The prognostic performance was assessed using the area under receiver operator characteristic curve (AUC).

Stability

Stability of feature selection and classification methods was assessed using previously defined stability measures (stability and RSD) (30). Stability measures were computed using the training cohort (HN1), and results were reported as median ± SD of 100 bootstrap iterations. It should be noted that in order to compute classifier stability, we first selected 30 representative features using MRMR. These selected features were then used as classifier input while computing classifier stability (RSD). We used MRMR because it showed the highest prognostic performance among all feature selection methods.

Prognostic Performance and Stability

As similar to Ref. (30), we used the median values of AUC and stability as thresholds to categorize the feature selection and classification methods into low or high performance (stability) groups. We created two rank lists based on AUC and stability and cited the methods as highly accurate and reliable, which ranked in the top half (greater than or equal to median value) in both ranked lists. Feature selection methods with stability $\geq$0.66 (median stability of all feature selection methods) and AUC $\geq$ 0.61 (median AUC of all feature selection methods) are considered as highly reliable and accurate methods. Similarly, classification methods with RSD $\leq$ 11.4 (median RSD of all classifiers) and AUC $\geq$ 0.61 (median AUC of all classifiers) are considered as highly reliable and accurate ones.

Experimental Factors Affecting Radiomics-Based Survival Prediction

Multifactor analysis of variance (ANOVA) was used to assess the variability in survival prediction. Three experimental factors were considered for the variability analysis: feature selection method, classification method, and the number of selected features. In order to compare the variability contributed by each factor and their interactions, the estimated variance components were divided by the total variance.

Comparison with the NSCLC Cohort Study

The results of this study were relatively compared with the previously published study of NSCLC radiomic cohorts (30). For both NSCLC and HNSCC studies, all methods were categorized into two groups: low (less than threshold) or high (greater than threshold). This grouping was carried out using the corresponding threshold values (median AUC and median stability). A method was considered consistent, if it belonged to the same group (high or low) in both studies. It should be noted that one feature selection method and one classification method was removed from the analysis for the HNSCC study, and therefore they were also not considered while deciding thresholds for NSCLC study. All values of AUC and stability along with the group information are reported in **Table 2** (feature selection methods) and **Table 3** (classification methods).

All analysis was done using R software (R Core Team, Vienna, Austria) version 3.1.2 and Matlab R2012b (The Mathworks, Natick, MA, USA) with Windows 7.

RESULTS

To assess different machine-learning methods for radiomic survival models of head and neck cancer patients, we extracted 440 radiomic features from the segmented tumor regions of two independent HNSCC cohorts. Cohort HN1 (n = 101 patients) was used for feature selection and classification training, whereas

TABLE 2 | Table describing the representative AUC and stability of feature selection methods.

Feature selection method	AUC (HNSCC)	AUC (NSCLC)	Stability (HNSCC)	Stability (NSCLC)
RELF	0.62 ± 0.09 (High)	0.61 ± 0.04 (High)	0.63 ± 0.12 (Low)	0.91 ± 0.05 (High)
FSCR	0.63 ± 0.08 (High)	0.62 ± 0.04 (High)	0.51 ± 0.13 (Low)	0.78 ± 0.08 (High)
GINI	0.58 ± 0.07 (Low)	0.62 ± 0.04 (High)	0.66 ± 0.11 (High)	0.68 ± 0.10 (Low)
JMI	0.59 ± 0.07 (Low)	0.61 ± 0.04 (High)	0.67 ± 0.05 (High)	0.68 ± 0.05 (Low)
CIFE	0.68 ± 0.08 (High)	0.60 ± 0.03 (Low)	0.7 ± 0.04 (High)	0.69 ± 0.05 (Low)
DISR	0.56 ± 0.06 (Low)	0.62 ± 0.05 (High)	0.65 ± 0.08 (Low)	0.69 ± 0.05 (Low)
MIM	0.61 ± 0.08 (High)	0.61 ± 0.04 (High)	0.64 ± 0.1 (Low)	0.94 ± 0.02 (High)
CMIM	0.6 ± 0.07 (Low)	0.62 ± 0.04 (High)	0.71 ± 0.04 (High)	0.73 ± 0.04 (Low)
ICAP	0.59 ± 0.07 (Low)	0.61 ± 0.03 (High)	0.71 ± 0.03 (High)	0.72 ± 0.04 (Low)
TSCR	0.62 ± 0.07 (High)	0.61 ± 0.02 (High)	0.54 ± 0.01 (Low)	0.78 ± 0.12 (High)
MRMR	0.69 ± 0.07 (High)	0.63 ± 0.06 (High)	0.66 ± 0.03 (High)	0.74 ± 0.03 (High)
MIFS	0.66 ± 0.07 (High)	0.63 ± 0.06 (High)	0.69 ± 0.04 (High)	0.8 ± 0.03 (High)
WLCX	0.55 ± 0.06 (Low)	0.65 ± 0.02 (High)	0.71 ± 0.06 (High)	0.84 ± 0.05 (High)

HNSCC thresholds: AUC = 0.61, stability = 0.66; NSCLC thresholds: AUC = 0.61, stability = 0.74.

TABLE 3 | Table describing the representative AUC and stability of classification methods.

Classification method	AUC (HNSCC)	AUC (NSCLC)	RSD% (HNSCC)	RSD% (NSCLC)
Nnet	0.59 ± 0.07 (Low)	0.57 ± 0.04 (Low)	11.54 (Low)	6.41 (Low)
DT	0.56 ± 0.05 (Low)	0.54 ± 0.04 (Low)	11.4 (High)	7.89 (Low)
BST	0.56 ± 0.07 (Low)	0.58 ± 0.04 (Low)	11.28 (High)	8.23 (Low)
BY	0.67 ± 0.06 (High)	0.64 ± 0.05 (High)	11.28 (High)	0.86 (High)
BAG	0.55 ± 0.06 (Low)	0.64 ± 0.03 (High)	9.27 (High)	5.56 (High)
RF	0.61 ± 0.06 (High)	0.66 ± 0.03 (High)	7.36 (High)	3.52 (High)
MARS	0.58 ± 0.05 (Low)	0.61 ± 0.03 (High)	12.47 (Low)	6.98 (Low)
SVM	0.64 ± 0.09 (High)	0.61 ± 0.03 (High)	12.69 (Low)	6.39 (Low)
NN	0.62 ± 0.05 (High)	0.61 ± 0.02 (High)	10.52 (High)	4.08 (High)
GLM	0.72 ± 0.08 (High)	0.63 ± 0.02 (High)	11.78 (Low)	2.19 (High)
PLSR	0.73 ± 0.07 (High)	0.63 ± 0.02 (High)	12.75 (Low)	2.24 (High)

HNSCC thresholds: AUC = 0.61, RSD% = 11.4; NSCLC thresholds: AUC = 0.61, RSD% = 5.56.

the prediction performance was assessed using the validation cohort HN2 ($n = 95$ patients) (see **Figure 1**).

Prognostic Performance

We used area under receiver operating characteristics curve (AUC) to quantify the prognostic performance of different feature selection and classification methods. **Figure 2** depicts the performance of feature selection (in rows) and classification methods (in columns) using the 30 top ranked features after applying feature selection. A median AUC of all 13 feature selection AUC values was used as the representative AUC of a classifier. Similarly, for each feature selection method, a median of 11 classification AUCs was used as the representative AUC. These representative AUC values for the feature selection and classification methods are given in **Tables 2** and **3**, respectively.

Feature selection methods, MRMR (AUC: 0.69 ± 0.07) and MIFS (AUC: 0.66 ± 0.07) showed high prognostic performance, whereas methods WLCX (AUC: 0.55 ± 0.06) and DISR (AUC: 0.56 ± 0.06) had lowest median AUCs.

In the case of classification methods, GLM (AUC: 0.72 ± 0.08) (median ± SD) and PLSR (AUC: 0.73 ± 0.07) had highest prognostic performance, whereas BAG (AUC: 0.55 ± 0.06), DT (AUC: 0.56 ± 0.05), and BST (AUC: 0.56 ± 0.07) showed lower AUC values. We repeated the above experiment by varying the number of selected features (range 5–50). Results with respect to 10, 20, 40, and 50 representative (top ranked) features are reported in supplement Figures S1–S4 in Supplementary Material. In addition, median AUC values over each of the experimental factors (feature selection methods, classification methods, and number of selected features) are depicted by the heatmaps in supplement Figures S5–S7 in Supplementary Material. Here as well, GLM and PLSR (classifiers) and MRMR and MIFS (feature selection methods) showed highest median AUCs in majority of cases.

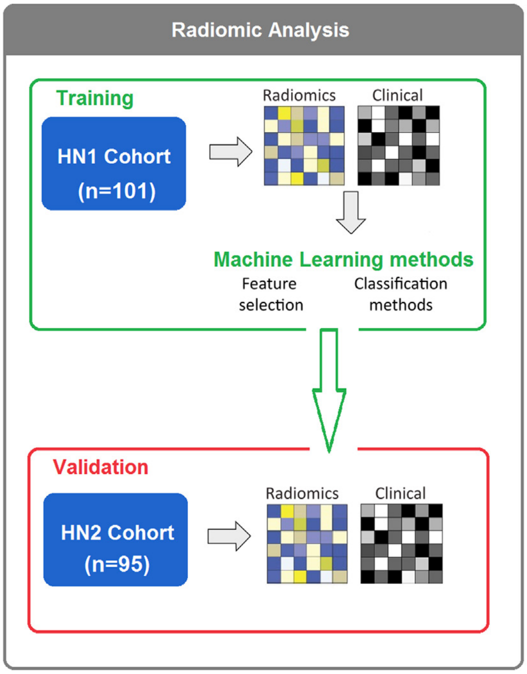


FIGURE 1 | In total, 196 HNSCC patients were considered. Four hundred forty radiomic features were extracted from the segmented tumor regions of the CT images. Feature selection and classification training were done using the training cohort HN1 ($n = 101$), whereas HN2 ($n = 95$) cohort was used as a validation cohort.

Stability

To assess the stability of feature selection methods against data perturbation, we used the hard data perturbation setting (37). We observed that WLCX (stability = 0.71 ± 0.06) (median ± SD), ICAP (stability = 0.71 ± 0.03), and CMIM (stability = 0.71 ± 0.04) showed high stability against data perturbation, whereas FSCR (stability = 0.51 ± 0.13) and TSCR (stability = 0.54 ± 0.10) had lower stability (**Table 2**).

In order to assess the stability of a classifier, we used the relative standard deviation (RSD) and a bootstrap approach. We observed that RF (RSD = 7.36%) and BAG (9.27%) were relatively more stable classification methods. PLSR (RSD = 12.75%) and SVM (RSD = 12.69%) showed higher RSD, which indicated lower stability of these methods. RSD (%) values corresponding to all 11 classifiers are reported in **Table 3**.

Prognostic Performance and Stability

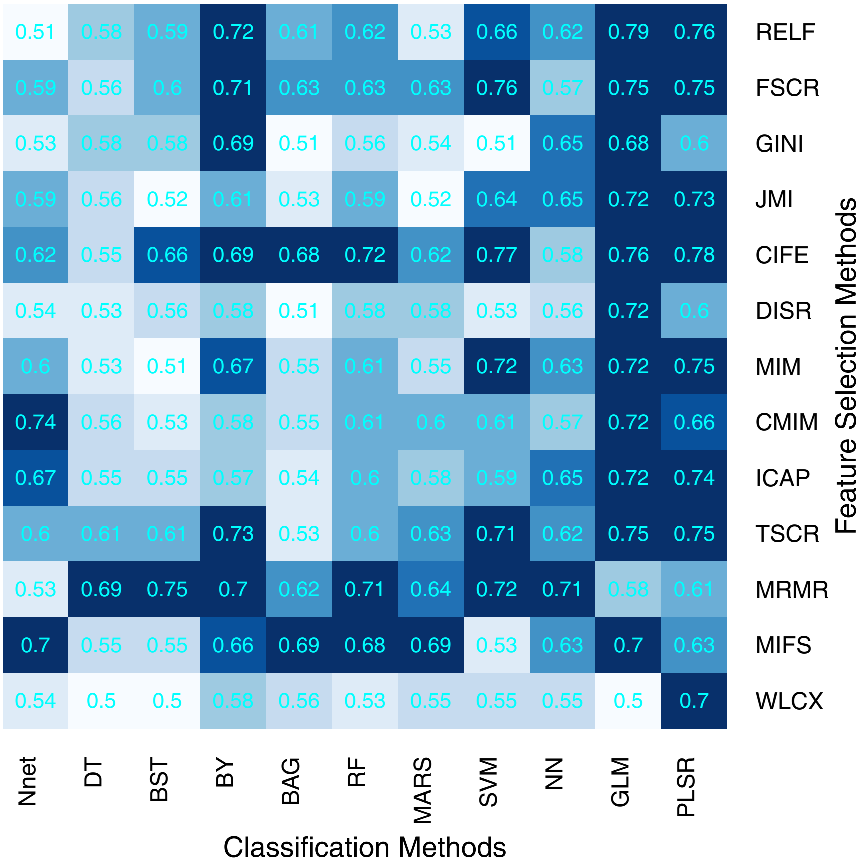
Scatterplots in **Figure 3** display the stability and prognostic performance of different feature selection and classification methods. It can be observed that the feature selection methods MRMR (AUC = 0.69 ± 0.07, stability = 0.66 ± 0.03), MIFS (AUC = 0.66 ± 0.07, stability = 0.69 ± 0.04), and CIFE (AUC = 0.68 ± 0.08, stability = 0.7 ± 0.04) showed higher prognostic performance and stability than the corresponding median values across all feature selection methods (AUC = 0.61, stability = 0.66). Similarly for classification methods, the stability and prognostic performance of RF (AUC = 0.61 ± 0.06, RSD = 7.36%), NN (AUC = 0.62 ± 0.05, RSD = 10.52%), and BY (AUC = 0.67 ± 0.06, RSD = 11.28%) were better than the corresponding median values (RSD = 11.4%, AUC = 0.61).

Experimental Factors Affecting Radiomics-Based Survival Prediction

To quantify the effects of the three experimental factors (feature selection methods, classification methods, and the number of selected features), we performed multifactor ANOVA on AUC scores. We observed that all three experimental parameters are the significant factors affecting the prognostic performance (**Figure 4**). Classification method was the most dominant source of variability as it explained 29.02% of the total variance in AUC scores. Feature selection methods accounted for the 14.02%, whereas interaction of classifier and feature selection explained 16.59% of the total variance. Size of the selected (representative) feature subset only shared 1.22% of the total variance (**Figure 4**).

Comparison with the NSCLC Cohorts

We compared the obtained results related to feature selection and classification methods to the previously published study of NSCLC cohorts (30) (see **Tables 2** and **3**).



	Nnet	DT	BST	BY	BAG	RF	MARS	SVM	NN	GLM	PLSR
RELF	0.51	0.58	0.59	0.72	0.61	0.62	0.53	0.66	0.62	0.79	0.76
FSCR	0.59	0.56	0.6	0.71	0.63	0.63	0.63	0.76	0.57	0.75	0.75
GINI	0.53	0.58	0.58	0.69	0.51	0.56	0.54	0.51	0.65	0.68	0.6
JMI	0.59	0.56	0.52	0.61	0.53	0.59	0.52	0.64	0.65	0.72	0.73
CIFE	0.62	0.55	0.66	0.69	0.68	0.72	0.62	0.77	0.58	0.76	0.78
DISR	0.54	0.53	0.56	0.58	0.51	0.58	0.58	0.53	0.56	0.72	0.6
MIM	0.6	0.53	0.51	0.67	0.55	0.61	0.55	0.72	0.63	0.72	0.75
CMIM	0.74	0.56	0.53	0.58	0.55	0.61	0.6	0.61	0.57	0.72	0.66
ICAP	0.67	0.55	0.55	0.57	0.54	0.6	0.58	0.59	0.65	0.72	0.74
TSCR	0.6	0.61	0.61	0.73	0.53	0.6	0.63	0.71	0.62	0.75	0.75
MRMR	0.53	0.69	0.75	0.7	0.62	0.71	0.64	0.72	0.71	0.58	0.61
MIFS	0.7	0.55	0.55	0.66	0.69	0.68	0.69	0.53	0.63	0.7	0.63
WLCX	0.54	0.5	0.5	0.58	0.56	0.53	0.55	0.55	0.55	0.5	0.7

FIGURE 2 | Heatmap depicting the prognostic performance (AUC) of feature selection (in rows) and classification (in columns) methods. It can be observed that PLSR and GLM classification methods and feature selection methods MRMR and MIFS shows relatively high prognostic performance in many cases.

Two feature selection methods MRMR and MIFS displayed high prognostic performance and stability across both the cancer types. Methods RELF and MIM showed high prognostic power in both the cancer types. However, they had marginally low stability for HNSCC cohorts. Interestingly, WLCX had highest prognostic performance for NSCLC cohorts, whereas it showed lowest prognostic performance for HNSCC cohorts. However, in both the cancer types WLCX displayed high stability against data perturbation (see **Table 2**).

Three classification methods BY, RF, and NN showed high prognostic performance and stability across both cancer types. PLSR and GLM had high prognostic performance in both cancer types, but these two methods showed low stability for HNSCC cohorts. Classifier BAG displayed lowest prognostic power in HNSCC radiomic cohorts, whereas it had the second highest performance in NSCLC cohorts. It should be noted that the stability of BAG was high in both cancer types (see **Table 3**).

DISCUSSION

The applications of medical imaging in cancer diagnostics and treatment planning have expanded greatly over the time. Moreover, developments in imaging technologies and computational approaches have led the emergence of "radiomics," which is a high-throughput medical image data mining field (1). Radiomics is a non-invasive and cost-effective medical informatics approach, which provides unprecedented opportunities for improving clinical decision support (2). Hence, in the context of radiomics, medical imaging is expected to have a more central role in cancer care (1, 2, 26). With increasing cohort sizes and expanding feature dimensions, radiomics targets a large pool of medical imaging data ("Big data"). Automated, reliable, and efficient methods are desired to extract and mine the most relevant information from these large radiomic cohorts. A recently published study (30) has articulated the scope and applicability of different machine-learning methods in two independent radiomic cohorts of patients with NSCLC. They proposed an unbiased framework to compare the prognostic performance and stability of different machine-learning methods. It was recommended that these different machine-learning methods should be further evaluated in different cancer types and with respect to different radiomic cohorts (30). Furthermore, it has been previously shown that the grouping and prognostic characteristics of radiomic features are cancer specific (16). Therefore, the primary objective of our study was to assess the state of the art machine-learning methods in two independent HNSCC cohorts.

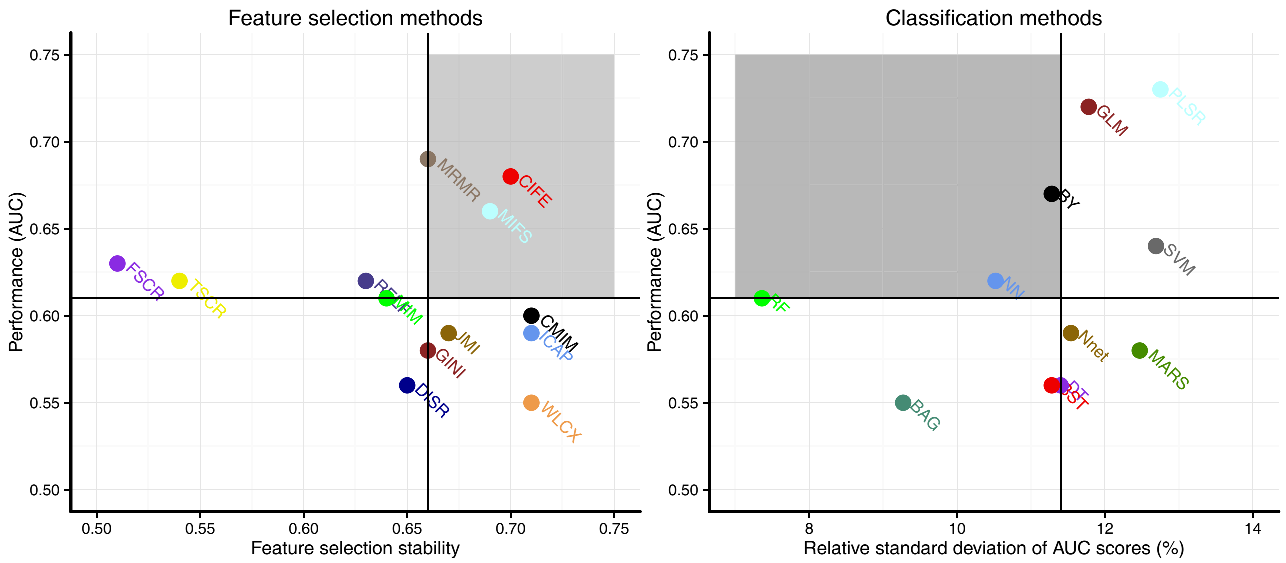


FIGURE 3 | Scatterplots of stability and prognostic performance (AUC) for feature selection (Left) and classification methods (right). Feature selection methods having stability ≥0.66 (median stability) and AUC ≥ 0.61 (median AUC) are considered as highly reliable and prognostic methods. Similarly, classification methods having RSD ≤ 11.4 (median RSD) and AUC ≥ 0.61 (median AUC) are considered as highly reliable and accurate ones. Highly reliable and prognostic methods are displayed in a gray square region.

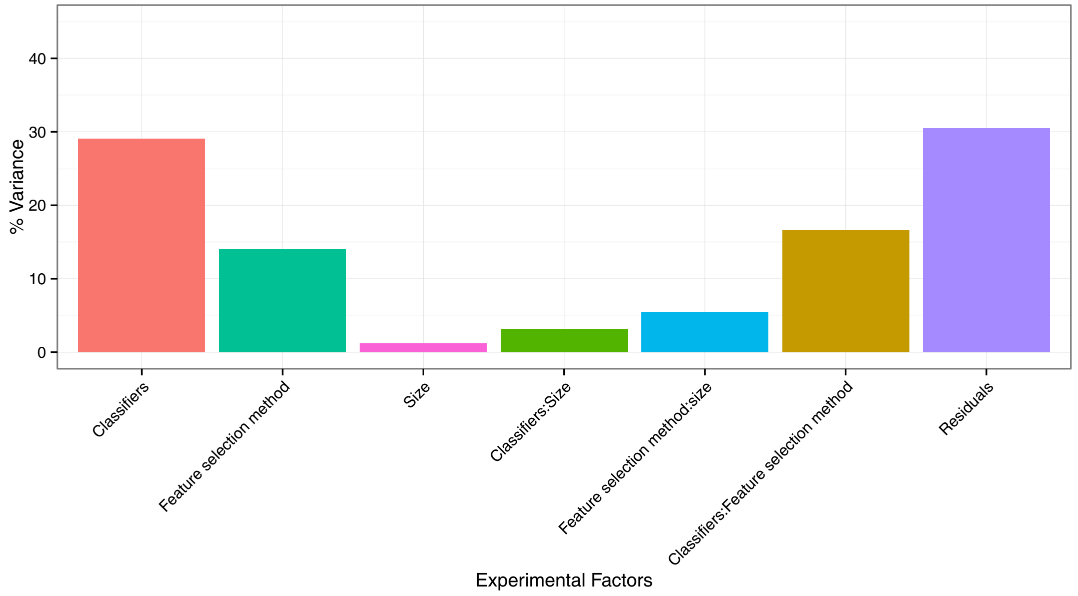


FIGURE 4 | Variation of AUC explained by the experimental factors and their interactions. It can be observed that classification method was the most dominant source of variation in prediction score. Size of the selected (representative) feature subset shared the least of the total variance.

Our analysis quantifies the prognostic power and stability of different machine-learning methods for the survival prediction of head and neck cancer patients. Depending on the requirement, one may prefer higher prognostic power or stability and choose the methods accordingly. Considering the stability and prognostic performance together, three feature selection methods MRMR, MIFS, and CIFE and three classification methods BY, RF, and NN should be preferred for head and neck radiomic analyses as they displayed relatively higher prognostic power and stability than other methods.

Assessing the variability in prediction performance by multifactor ANOVA, we found that the classification method is the most dominant source of variation in prediction performance, and hence it should be chosen carefully. Size of the selected feature subset contributed the least in the AUC variation.

Comparing the results of this study with the previously published study of NSCLC, we observed that except BAG and MARS, all classifiers showed consistency in their prognostic performance across the two cancer types. However, it should be noted that the AUC values for MARS were quite close to the threshold (median AUC). In the case of feature selection methods, WLCX showed no prediction consistency across the two cancer types. Besides that, methods CIFE, DISR, GINI, CMIM, and ICAP also showed no consistency in the grouping based on prognostic performance. However, it should be noted that for methods CMIM and ICAP, the AUC values did not differ much from the threshold (median AUC). As far as stability is concerned, except for classifiers PLSR, GLM, DT, and BST, all classifiers showed consistency in their stability-based grouping. For feature selection methods, only MRMR, MIFS, WLCX, and DISR showed consistent stability-based grouping across both the studies. It should be noted that for HNSCC cohorts, stability values for all the methods (classification and feature selection) were lower than the ones obtained in NSCLC cohorts. The intuitive explanation for this could be the smaller cohort size. The training cohort in NSCLC study (30) had almost three times more samples than the training cohort of our HNSCC study.

Considering the stability and prognostic performance together and comparing the results between the two cancer types (HNSCC and NSCLC), we observed that the three classifiers BY, RF, and NN had high stability and high prognostic performance in both HNSCC and NSCLC studies. Similarly, two feature selection methods MRMR and MIFS showed consistently high values of AUC and stability in both cancer types. These results indicate that such methods should be first preferences for radiomics-based prognostic analyses due to their consistency. A note of caution: different methods are categorized into high/low (prediction performance and stability) group based on simple thresholds (median AUC and median stability). These thresholds are no gold standard and they are only used for comparing the performances of different machine-learning methods in a relative manner. It can be observed from the results [**Figure 3**; Figure 3 in Ref. (30)] that some of the so-called "not consistent" methods are quite close to the thresholds, and they should not be neglected completely. Further validation of these methods with different clinical outcomes, different imaging modalities, and also different radiomic cohorts could provide better insights about their applicability.

Results related to the variability of AUC scores were comparable in both the cancer types as in both studies, classification method contributed highest and size contributed the least in the performance variance. Interestingly, for HNSCC cohorts, feature selection method contributed almost two times more in the AUC variation than in the case of NSCLC study (30).

As mentioned previously (30), the machine-learning methods used in this analysis were chosen because of their simplicity, efficiency, and popularity in the literature. Furthermore, an interesting discussion about the publicly available implementation tools and the used parameter configurations were presented before (30).

It has been shown that statistical models based on patient's tumor and treatment characteristics provide significantly better predictions/prognosis than human expert (38). Therefore, radiomics-based machine-learning models could be vital for clinical decision support. However, there are few inherent challenges with radiomics. Several studies have investigated the variability of radiomic features with respect to different imaging scanners (39, 40), tumor delineation methods (13, 41), reconstruction methods (42), discretization (43), etc. These different sources of variability need to be considered for radiomics-based analyses. For example, normalizing or standardizing the features could help in reducing batch effects. Furthermore, the performance of prediction models should be tested in independent validation cohorts. As far as our study is concerned, we standardized all features using *Z*-score standardization. Moreover, we used an independent validation cohort to assess the prediction performance of different machine-learning methods.

The potential clinical utility of radiomics-based prognostic models has been highlighted before (2). With the expanding radiomic cohorts and feature dimensions, as well as by integrating different biological and clinical information together with radiomics, higher prognostic performance could be achieved. In this regard, our studies could be an important reference as we compared a large panel of machine-learning methods across two different cancer types. The prognostic power and stability of different machine-learning methods were compared using four independent radiomic cohorts. Such a comparative investigation could help in identifying the optimal and reliable machine-learning methods for radiomics-based prognostic analyses, which overall could broaden the scope of radiomics in cancer care.

AUTHOR CONTRIBUTIONS

HA, CP, and PG conceived of the project, analyzed the data, and wrote the paper. DR, MR, and PL provided expert guidance, data, or analysis tools and reviewed the manuscript.

ACKNOWLEDGMENTS

Authors acknowledge financial support from the National Institute of Health (NIH-USA U24CA194354 and NIH-USA U01CA190234). This research is also supported by the Dutch technology Foundation STW (grant no. 10696 DuCAT). Authors also acknowledge financial support from EU 7th framework program (ARTFORCE – no. 257144, REQUITE – no. 601826), SME Phase 2 (EU proposal 673780 – RAIL), the European Program H2020-PHC-2015 (BD2decide, no. 210274050), Alpe d'HuZes-KWF (DESIGN).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fonc.2015.00272

REFERENCES

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) **48**(4):441–6. doi:10.1016/j.ejca.2011.11.036
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* (2014) **5**:4006. doi:10.1038/ncomms5006
3. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* (2013) **108**(3):479–85. doi:10.1038/bjc.2012.581
4. Ng CK, Pemberton HN, Reis-Filho JS. Breast cancer intratumor genetic heterogeneity: causes and implications. *Expert Rev Anticancer Ther* (2012) **12**(8):1021–32. doi:10.1586/era.12.85
5. Szerlip NJ, Pedraza A, Chakravarty D, Azim M, McGuire J, Fang Y, et al. Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proc Natl Acad Sci U S A* (2012) **109**(8):3041–6. doi:10.1073/pnas.1114033109
6. Longo DL. Tumor heterogeneity and personalized medicine. *N Engl J Med* (2012) **366**(10):956–7. doi:10.1056/NEJMe1200656
7. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PLoS One* (2014) **9**(10):e110300. doi:10.1371/journal.pone.0110300
8. Carvalho S, Leijenaar RT, Velazquez ER, Oberije C, Parmar C, Van Elmpt W, et al. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. *Acta Oncol* (2013) **52**(7):1398–404. doi:10.3109/0284186X.2013.812795
9. Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, et al. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* (2014) **273**(1):168–74. doi:10.1148/radiol.14131731
10. Hatt M, Tixier F, Le Rest CC, Pradier O, Visvikis D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging* (2013) **40**(11):1662–71. doi:10.1007/s00259-013-2486-8
11. Jain R, Poisson LM, Gutman D, Scarpace L, Hwang SN, Holder CA, et al. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* (2014) **272**(2):484–93. doi:10.1148/radiol.14131691
12. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol* (2013) **52**(7):1391–7. doi:10.3109/0284186X.2013.812798
13. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* (2014) **9**(7):e102107. doi:10.1371/journal.pone.0102107
14. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, Kry SF. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. *Med Phys* (2013) **40**(12):121916. doi:10.1118/1.4829514
15. Basu S, Hall LO, Goldgof DB, Gu Y, Kumar V, Choi J, et al., editors. Developing a classifier model for lung tumors in CT-scan images. *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference On*. IEEE (2011) (Anchorage, AK). p. 1306–12.
16. Parmar C, Leijenaar RT, Grossmann P, Velazquez ER, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep* (2015) **5**:11044. doi:10.1038/srep11044
17. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. *Radiology* (2013) **266**(1):326–36. doi:10.1148/radiol.12112428
18. Cistaro A, Quartuccio N, Mojtahedi A, Fania P, Filosso PL, Campenni A, et al. Prediction of 2 years-survival in patients with stage I and II non-small cell lung cancer utilizing 18F-FDG PET/CT SUV quantifica. *Radiol Oncol* (2013) **47**(3):219–23. doi:10.2478/raon-2013-0023
19. Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? *J Nucl Med* (2013) **54**(1):19–26. doi:10.2967/jnumed.112.107375
20. Hawkins SH, Korecki JN, Balagurunathan Y, Gu Y, Kumar V, Basu S, et al. *Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features*. (IEEE) (2014). p. 1418–26.
21. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* (2010) **10**(1):137. doi:10.1102/1470-7330.2010.0021
22. Coroller TP, Grossmann P, Hou Y, Velazquez ER, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* (2015) **114**(3):345–50. doi:10.1016/j.radonc.2015.02.015
23. Pickles MD, Manton DJ, Lowry M, Turnbull LW. Prognostic value of pre-treatment DCE-MRI parameters in predicting disease free and overall survival for breast cancer patients undergoing neoadjuvant chemotherapy. *Eur J Radiol* (2009) **71**(3):498–505. doi:10.1016/j.ejrad.2008.05.007
24. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* (2007) **25**(6):675–80. doi:10.1038/nbt1306
25. Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder CA, Gutman D, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. *J Neuroradiol* (2015) **42**:212–21. doi:10.1016/j.neurad.2014.02.006
26. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology – multifactorial decision support systems. *Nat Rev Clin Oncol* (2013) **10**(1):27–40. doi:10.1038/nrclinonc.2012.196
27. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging* (2012) **30**(9):1234–48. doi:10.1016/j.mri.2012.06.010
28. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. Cambridge, MA: MIT Press (2012).
29. Guyon I, André E. An introduction to variable and feature selection. *J Mach Learn Res* (2003) **3**:1157–82.
30. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* (2015) **5**:13087. doi:10.1038/srep13087
31. Brown G, Pocock A, Zhao M-J, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J Mach Learn Res* (2012) **13**(1):27–66.
32. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H. *Advancing Feature Selection Research. ASU Feature Selection Repository*. Tempe, AZ: School of Computing, Informatics, and Decision Systems Engineering, Arizona State University (2010).
33. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* (2014) **15**(1):3133–81.
34. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern* (1973) **3**(6):610–21.
35. Galloway MM. Texture analysis using gray level run lengths. *Comput Graphics Image Process* (1975) **4**(2):172–9. doi:10.1016/S0146-664X(75)80008-6
36. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* (2008) **28**(5):1–26. doi:10.18637/jss.v028.i05
37. Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One* (2011) **6**(12):e28210. doi:10.1371/journal.pone.0028210
38. Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother Oncol* (2014) **112**(1):37–43. doi:10.1016/j.radonc.2014.04.012
39. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol* (2015) **50**(11):757–65. doi:10.1097/RLI.0000000000000180
40. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring variability in CT characterization of tumors: a preliminary phantom study. *Transl Oncol* (2014) **7**(1):88–93. doi:10.1593/tlo.13865

41. Velazquez ER, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep* (2013) **3**:3529. doi:10.1038/srep03529

42. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol* (2010) **49**(7):1012–6. doi:10.3109/0284186X.2010.498437

43. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep* (2015) **5**:11075. doi:10.1038/srep11075

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.