Title:

Brain metastases from NSCLC treated with stereotactic radiotherapy: prediction mismatch between two different radiomic platforms

Authors:

Gianluca Carloni[a*,b,c,1], Cristina Garibaldi[d,1], Giulia Marvaso[a,e], Stefania Volpe[a,e], Mattia Zaffaroni[a,+], Matteo Pepa[a], Lars Johannes Isaksson[a,e], Francesca Colombo[a,e], Stefano Durante[a], Giuliana Lo Presti[f*], Sara Raimondi[f], Lorenzo Spaggiari[e,g], Filippo De Marinis[h], Gaia Piperno[i], Sabrina Vigorito[i], Sara Gandini[f], Marta Cremonesi[d], Vincenzo Positano[c,l,2], and Barbara Alicja Jereczek-Fossa[a,e,2]

a *Division of Radiation Oncology, IEO, European Institute of Oncology, IRCCS, Milan, Italy*
b *Institute of Information Science and Technologies, National Research Council, Pisa, Italy*
c *Department of Information Engineering, University of Pisa, Pisa, Italy*
d *Unit of Radiation Research, IEO, European Institute of Oncology, IRCCS, Milan, Italy*
e *Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy*
f *Department of Experimental Oncology, IEO, European Institute of Oncology, IRCCS, Milan, Italy*
g *Department of Thoracic Surgery, IEO, European Institute of Oncology IRCCS, Milan, Italy*
h *Division of Thoracic Oncology, IEO, European Institute of Oncology, IRCCS, Milan, Italy*
i *Unit of Medical Physics, IEO, European Institute of Oncology, IRCCS, Milan, Italy*
l *Gabriele Monasterio Foundation, Pisa, Italy*

[*]*Affiliation at the time of the study*
[1]*Co-first authors*
[2]*Co-last authors*
[+]*Corresponding author:* Mattia Zaffaroni, MSc, Division of Radiation Oncology IEO, European Institute of Oncology IRCCS, Via Ripamonti, 435, 20141, Milano, Italy, mattia.zaffaroni@ieo.it

List of Abbreviations

ALK, Anaplastic lymphoma kinase

BED, Biologically Effective Dose

BM, Brain Metastasis

C-index, Concordance index

CR, Complete Response

CT, Computed Tomography

DP, Distant Progression

EGFR, Epidermal growth factor receptor

EQD2, Equivalent dose in 2Gy fractions

GLCM, Gray Level Co-Occurrence Matrix

GLDZM, Gray Level Distance Zone Matrix

GLRLM, Gray Level Run Length Matrix

GLSZM, Gray Level Size Zone Matrix

HR, Hazard Ratio

IBSI, Imaging Biomarker Standardization Initiative

IEO, Istituto Europeo di Oncologia (European Institute of Oncology) IRCCS, Milan, Italy

KM, Kaplan-Meier

KPS, Karnofsky Performance Status

LASSO, Least Absolute Shrinkage and Selection Operator

LC, Local Control

LoG, Laplacian of Gaussian

MRI, Magnetic Resonance Imaging

NGLDM, Neighbourhood Gray Level Dependence Matrix

NGTDM, Neighbouring Gray Tone Difference Matrix

NSCLC, Non-Small Cell Lung Cancer

OS, Overall Survival

PD, Progression Disease

PR, Partial Response

PyR, PyRadiomics

RS, Radiomic Score

RT, Radiotherapy

RTSS, Radiation Therapy Structure Sets

SD, Stable Disease

SR, SOPHiA Radiomics

SRS, Stereotactic Radiosurgery

T1-w, T1-weighted

Abstract

Background and purpose. Radiomics enables the mining of quantitative features from medical images. The influence of the radiomic feature extraction software on the final performance of models is still a poorly understood topic. This study aimed to investigate the ability of radiomic features extracted by two different radiomic platforms to predict clinical outcomes in patients treated with radiosurgery for brain metastases from non-small cell lung cancer. We developed models integrating pre-treatment magnetic resonance imaging (MRI)-derived radiomic features and clinical data.

Materials and Methods. Pre-radiotherapy gadolinium enhanced axial T1-weighted MRI scans were used. MRI images were re-sampled, intensity-shifted, and histogram-matched before radiomic extraction by means of two different platforms (PyRadiomics and SOPHiA Radiomics). We adopted LASSO Cox regression models for multivariable analyses by creating radiomic, clinical, and combined models using three survival clinical endpoints (local control, distant progression, and overall survival). The statistical analysis was repeated 50 times with different random seeds and the median concordance index was used as performance metric of the models.

Results. We analysed 276 metastases from 148 patients. The use of the two platforms resulted in differences in both the quality and the number of extractable features. That led to mismatches in terms of end-to-end performance, statistical significance of radiomic scores, and clinical covariates found significant in combined models.

Conclusion. This study shed new light on how extracting radiomic features from the same images using two different platforms could yield several discrepancies. That may lead to acute consequences on drawing conclusions, comparing results across the literature, and translating radiomics into clinical practice.

1. Introduction

In the attempt to seek novel non-invasive strategies to characterise solid tumours and their response to treatment, radiomics has been gaining interest due to the growing availability of high-performance computing capabilities [1]. It consists in the extraction of hundreds to thousands quantitative imaging features from medical images, usually beyond human perception. Using advanced statistical methods, subsets of such features can be used to generate mathematical models representing distinctive signatures related to the underlying tumour biology, potentially helpful for the assessment of prognosis or treatment response from a precision-medicine viewpoint.

In recent years, potential obstacles to the translation of radiomics into clinical practice have been documented. Remarkable examples show that computed values of radiomic features can be severely biased by the acquisition and reconstruction settings of the scanner [2] [3] [4], or by inter-operator variability in lesion segmentation [5] [6].

In the radiotherapy (RT) setting, radiomics has been applied to predict treatment outcomes of several districts, including brain tumours, prostate cancer, oesophageal cancer, lung cancer, sarcoma, and rectal cancer [7] [8] [9] [10] [11].

Regarding brain metastases (BMs), radiomics is mainly based on the analysis of structural MRI data. The application of radiomics in BMs has encompassed several outcomes, resulting useful to differentiate treatment-related changes from BMs after RT [12] [13], to predict BMs origin [14] [15], to differentiate BMs from other malignancies [16], and to assess treatment response [17]. Particularly, an increasing number of works regarding radiomics on BMs from non-small cell lung cancer (NSCLC) is being reported in the literature [18] [19] [20].

A central component of the radiomic workflow, not fully explored yet, is the selection of the software platform for features calculation. In fact, several platforms with different characteristics (Imaging Biomarker Standardization Initiative (IBSI) compliance [21], open-source nature, documented mathematical equations, cost, etc.) have been developed over the last 7 years, each coming with similarities with others but also with ad-hoc peculiarities that might impact performance and prevent results comparison. A growing number of evidence in literature reports features' values variability across different packages, on either patient datasets [22] [23] [24], or phantoms [25].

The aim of this study is to investigate the ability of radiomic features extracted from two different radiomic platforms to predict clinical outcomes in patients treated with stereotactic radiosurgery (SRS) for BMs from NSCLC, by developing models integrating pre-treatment MRI-derived radiomic features and clinical data.

2. Materials and Methods

2.1. Patients

The study population was retrospectively selected from a database of patients with synchronous and metachronous BMs from NSCLC treated with SRS by CyberKnife® (Accuray Inc., Sunnyvale, CA) and concomitant pre- or post-systemic therapy at the European Institute of Oncology IRCCS, Milan, Italy (IEO).

Eligibility criteria for SRS included: age ≥18 years, histologically proven NSCLC and mutational burden status, Karnofsky Performance Status (KPS) > 60, radiologically proven BMs with pre-treatment images, not previous RT or surgical treatment for BMs before the first SRS and written informed consent for research and training purposes. The approval for the use of medical data in the study was provided by the institutional review board.

Clinical, mutational status (presence/absence of epidermal growth factor receptor (EGFR) mutation and/or anaplastic lymphoma kinase (ALK)-rearrange) and dosimetry data were collected for each patient. In particular, the considered records were: primary tumour data (date of diagnosis, T and N stage, histology, treatment strategy, mutational status), local outcome data (maximum volumetric response to RT, date of maximum response, progressive disease, therapy of progressive disease) and radiation treatments parameters, such as number of fractions, dose per fraction, total dose, biologically effective dose (BED), and equivalent dose in 2 Gy fractions (EQD2). Further details are provided in Supplementary S.1. A summary of the patient characteristics is presented in Table 1.

Variable	Level	Overall (N=148)
Age at start of RT (median, min-max)		65 (28-87)
Sex	F M	62 (41.9%) 86 (58.1%)
KPS (median, min-max)		90 (60-100)
Histology	Adenocarcinoma Squamous cell carcinoma Sarcomatous carcinoma Neuroendocrine carcinoma Not otherwise specified (NOS)	128 (86.5%) 12 (8.1%) 3 (2.0%) 3 (2.0%) 2 (1.4%)
Mutations	No (no EGFR mut. nor ALK-rearrange) Yes (EGFR mut. and/or ALK-rearrange)	105 (71%) 43 (29%)
Stage at diagnosis	I-III IV	69 (46.6%) 79 (53.4%)
BMs at the onset	No (metachronous) Yes (synchronous)	93 (62.8%) 55 (37.2%)
Surgery on the primary	No Yes	91 (61.5%) 57 (38.5%)
RT on the primary	No Yes	119 (80.4%) 29 (19.6%)
Extracranial metastases at the onset	No Yes	96 (64.9%) 52 (35.1%)
Number of lesions (median, min-max)		1 (1-6)
Number of lesions	1 2 3 4 6	77 (52.0%) 36 (24.3%) 20 (13.5%) 13 (8.8%) 2 (1.4%)
Total intracranial tumour volume cm^3 (median, min-max)		0.95 (0.03-26.9)
Prescription BED Gy (median, min-max)		53 (29-82)

Table 1. **Patient characteristics.** RT=radiotherapy, KPS=Karnofsky Performance Status, EGFR=epidermal growth factor receptor, ALK=anaplastic lymphoma kinase, BMs=brain metastases, BED=Biologically Effective Dose

2.2. Images acquisition and pre-processing

This study was conducted on pre-RT gadolinium enhanced axial T1-weighted (T1-w) MRI scans, registered to the simulation computed tomography (CT) for volume delineation. MRI scans and radiation therapy structure sets (RTSS) were collected for each patient from the treatment plans on Precision™ Treatment Planning System workstations (Accuray Inc., Sunnyvale, CA).

Prior to features extraction, the following pre-processing steps were implemented to ensure an accurate quantitative analysis (Supplementary S.2):

- Voxel re-sampling to correct for differences in pixel spacing and slice thicknesses
- Intensity shifting to have all images' minimum value equal to zero
- Intensity normalization to correct for scanner-dependent variations using the histogram matching normalization with (0th-50th-99th) percentiles of the intensities.

2.3. Radiomic feature extraction

The extraction of radiomic features was performed on the gross tumour volume of each BM of all patients using two platforms, PyRadiomics (PyR) [26] and SOPHiA Radiomics (SR) [27]. While PyR is a free, open-access package, SR is a commercial software whose licence was obtained under scientific agreements between IEO and the company.

To foster reproducibility and comparison, default extraction settings with PyR were utilised. In addition to the original images, Laplacian of Gaussian (LoG)–filtered images and Wavelet-transformed images were also used. The LoG filter with sigma values of 0.5, 2.75 and 5.0 mm, representing fine, medium, and coarse patterns, respectively, was used for image filtration. Wavelet-based texture features were generated exploiting Wavelet filtering, which yields eight decompositions per level, which are all the possible three-dimensional combinations of applying either a High or a Low pass filter. Only one level was selected. Features from all the available classes were extracted.

The following feature classes were enabled for the calculation of features in SR: gray level co-occurence matrix (GLCM), gray level distance zone matrix (GLDZM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), General, Intensity histogram, Local intensity, Morphological, neighborhood gray level dependence matrix (NGLDM), neighbouring gray tone difference matrix (NGTDM), Statistical (IBSI), Statistical (Original Data), Volume-Intensity Histogram.

2.4. Statistical analysis

All primary and secondary clinical endpoints are reported in Table 2.

Clinical Endpoint	Definition	Event definition	Time-to-event
Local Control	Maximum response of BMs to SRS. Complete response (CR), partial response (PR), stable disease (SD), progression disease (PD)	1: CR 0: PR, SD, or PD	Months from the start of SRS to date of maximum response, determined on follow-up MRI
Distant Progression	Any new BM developed outside the previous target volume	1: DP 0: no DP	Months from the start of SRS to the date of DP or last follow-up, whichever occurred first, determined on follow-up MRI
Overall Survival	Length of time from the start of treatment for BMs that patients are still alive	1: dead 0: alive	Months from the start of SRS to the date of death or last follow-up, whichever occurred first

Table 2. **Clinical endpoints.** BM = brain metastasis, SRS = stereotactic radiosurgery, MRI = magnetic resonance imaging

The curves of cumulative proportion of lesions with complete response (CR) and those with distant progression (DP), and the patients' overall survival (OS) curve were computed using the Kaplan-Meier (KM) [28] estimator. For OS prediction, a representative lesion for each multiple-lesion patient was considered. The lesion with worst maximum response was selected unless several of the patient's lesions had the same worst response. In this case the lesion whose volume was as close as possible to the median volume over all lesions was selected.

Three kinds of prognostic models were designed: radiomic, clinical and combined models. Radiomic models leveraged radiomic features extracted from pre-RT MRI scans to predict response to treatment or patients' survival. Clinical models investigated the same outcomes and were based on clinical, mutational and dosimetry features. Combined models were developed considering both kinds of features. Five different models were developed for each endpoint (Figure S.5): two radiomic (PyR, SR), one clinical, and two combined (PyR + clinical, SR + clinical). For all the models, estimated coefficients of their covariates were considered significant with a p-value < 0.05 and performance was evaluated with the resulting concordance index (C-index) [29]. To assess how the C-index results varied by assignment to the initial dataset, the analysis was independently repeated 50 times by using different random seeds in the data splitting process detailed below.

A common workflow was designed to develop radiomic models (Figure 1).

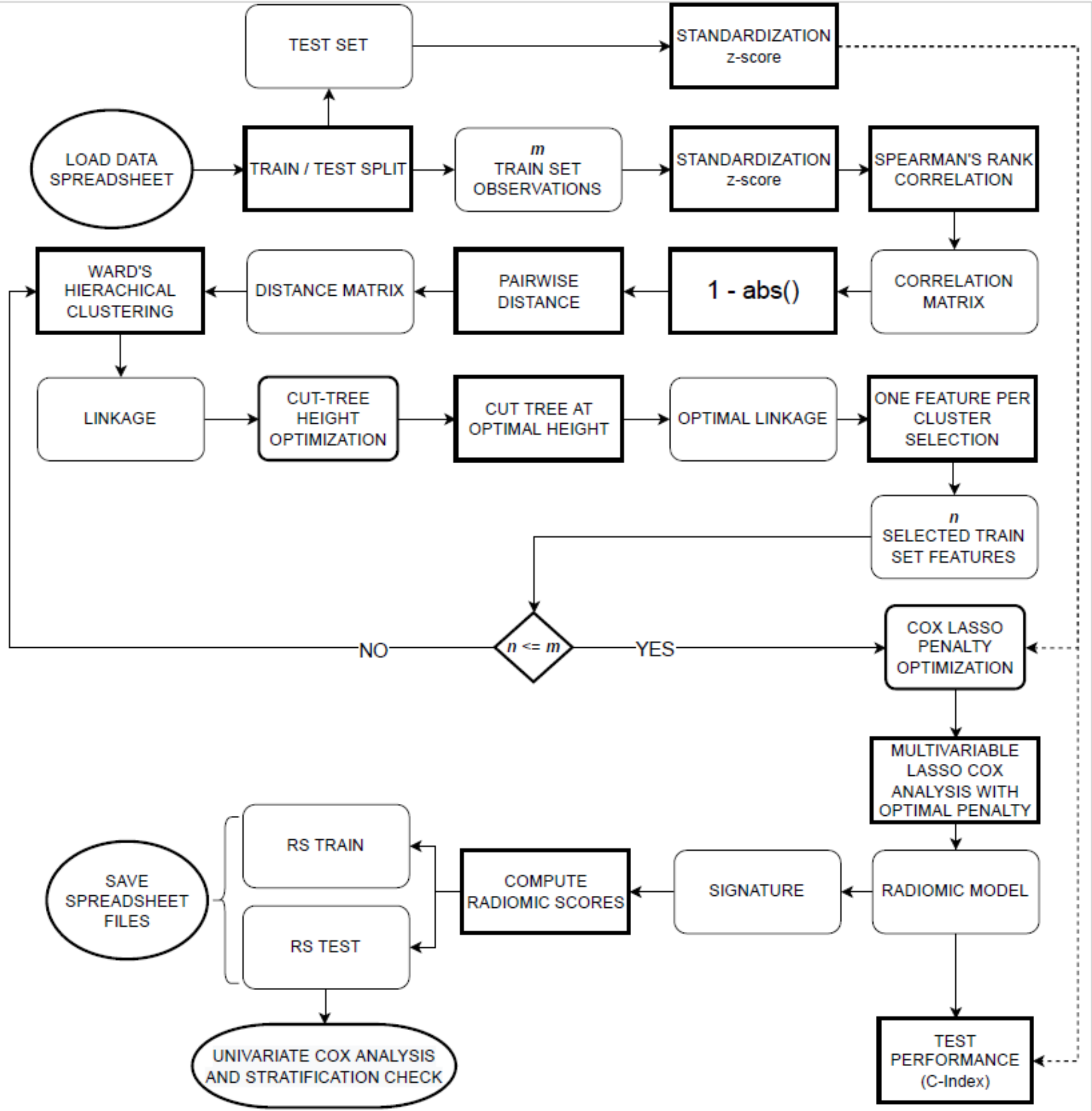
TEST SET
STANDARDIZATION z-score
LOAD DATA SPREADSHEET
TRAIN / TEST SPLIT
m TRAIN SET OBSERVATIONS
STANDARDIZATION z-score
SPEARMAN'S RANK CORRELATION
WARD'S HIERACHICAL CLUSTERING
DISTANCE MATRIX
PAIRWISE DISTANCE
1 - abs()
CORRELATION MATRIX
LINKAGE
CUT-TREE HEIGHT OPTIMIZATION
CUT TREE AT OPTIMAL HEIGHT
OPTIMAL LINKAGE
ONE FEATURE PER CLUSTER SELECTION
n SELECTED TRAIN SET FEATURES
NO
n <= m
YES
COX LASSO PENALTY OPTIMIZATION
MULTIVARIABLE LASSO COX ANALYSIS WITH OPTIMAL PENALTY
RS TRAIN
SAVE SPREADSHEET FILES
COMPUTE RADIOMIC SCORES
SIGNATURE
RADIOMIC MODEL
RS TEST
UNIVARIATE COX ANALYSIS AND STRATIFICATION CHECK
TEST PERFORMANCE (C-Index)

Figure 1. **Development pipeline for radiomic models**. *m* = number of train set observation, *abs()* = absolute value, *n* = number of selected train set features, C-index = concordance index, RS = radiomic score.

Initially, available lesions were shuffled and then split into train and test sets in a proportion of 2/3 and 1/3, respectively. The split was done in a stratified way so that, for each of the endpoints, the same proportion of events was maintained in train and test sets, and lesions from the same image (subject) belonged to the same subset. Subsequent operations were performed on the train set. Raw feature scores were converted in standard scores on the same scale; their pairwise correlation matrix was calculated, and from that the Euclidean pairwise distance matrix was obtained. Dimensionality reduction was pursued through feature clustering and elimination of highly correlated ones. A hierarchical clustering over the determined distance matrix was performed with Euclidean distance and Ward's method [30]. A hierarchical tree cut-height optimization process was carried out and a single feature per cluster was selected. To additionally reduce the number of features for prediction, the Least Absolute Shrinkage and Selection Operator (LASSO) [31] was integrated into the Cox's proportional hazard model [32], resulting in a penalised survival regression model. Several LASSO penalty values were tested during an optimization process (Supplementary S.4). The model was then fitted on train-set lesions, and every feature was reported along with the estimated coefficient, the exponential of that coefficient, i.e., the hazard ratio (HR), and the p-value of the estimate. The radiomic signature, as a list of coefficients estimated for each covariate, was then extracted. A Radiomic Score (RS), defined as the weighted sum of feature standardised values where the coefficients in the signature represented weights, was built both for train and test observations. A univariate Cox analysis was performed between the RS and the outcome on test-set observations. In addition, to evaluate prognostic significance of the RS, KM analysis was performed. Observations were divided into low-RS and high-RS groups, where the threshold was the median of RS values across all the train-set observations. The curves of low-RS and high-RS groups were considered significantly separated with a log-rank test p-value < 0.05.

To build the clinical models, a common workflow was followed (Figure S.11). A different set of clinical features was considered for LC/DP (lesion-based analyses) and OS (patient-based analysis); see Tables S.2 and S.3, respectively. For multivariable analysis, an unpenalised Cox model was fitted on train lesions considering the clinical features conditioned to the endpoint.

In the combined models, following the same workflow of the clinical ones (Figure S.12), the radiomic information is added as a score, RS, to all other features of the clinical models. Only significant features in the respective clinical model were included in the combined one.

In the end, for each model we computed the median, interquartile range (IQR), and notched boxplot of C-index values with 95% confidence interval (CI) calculated with 50 different random seeds. All models were developed in Python 3.8.5 environment; see Table S.1 for details.

2.4. Side Analyses

We carried out additional side analyses in this work. On the one hand, we repeated all the analyses by excluding LoG- and Wavelet-based features from the building of the models. This was done to prevent the inclusion of possible bias in results, since standards for such filters are still being developed [33]. On the other hand, we built models with the subset of features that are shared between both platforms. This was done to investigate whether possible differences would persist in this scenario. Even if features might present different naming conventions across the platforms, we identified and used 98 common features (see Supplementary S.4.g). We ran 50 executions of these analyses by using the same random seeds used in the primary analysis.

3. Results

Data of 198 patients with BMs from NSCLC treated between February 2012 and August 2018 with CyberKnife at IEO were retrospectively collected. Only data regarding the first treatment were considered. Fifty patients were excluded from the analysis due to a lack of a pre-RT T1-w MRI scan. We analysed 276 BMs from the resulting 148 patients who matched all the inclusion criteria.

Amongst the patients' MRI scans, 113 (76.4%) were acquired at IEO on a Siemens scanner, while 24 (16.2%) on a General Electric scanner, and 11 (7.4%) were acquired externally on various other scanners. After voxel re-sampling, all images shared common pixel spacing of 0.98x0.98 mm^2 and slice thickness of 1.25 mm. Images were histogram-matched to the triad (1, 64, 813) representative of the (0th-50th-99th) percentiles of the intensities of the entire dataset.

For each endpoint, the number of events and median time-to-event estimated from KM curves are reported in Table 3.

Endpoint	Number of events	Median time-to-event (months)
Local Control	53/276 (19.2%)	17.8
Distant Progression	143/276 (51.8%)	13.3
Overall Survival	71/148 (48.0%)	20.2

Table 3. **Number of events and median time-to-event for each endpoint.**

Among the main findings from the clinical and combined models, older patients at start of SRS and patients who undergo concomitant therapy were independently more likely to achieve CR. In addition, patients with stage IV cancer at the onset and, patients with BM at the onset resulted independently more prone to DP. Instead, concomitant therapy was a protective factor for DP compared to no therapy at all. Furthermore, patients with higher KPS and patients prescribed with higher BED were independently associated with extended OS. Conversely, a high total intracranial tumour volume was recognised as a risk factor for OS.

In Table 4 we present a summary of the distribution of C-index values on test set for each model across the 50 executions for the three different analyses: (i) different feature set, (ii) without LoG and Wavelet features, and (iii) common features only.

Endpoint	Model	Median C-index (Q1-Q3)		
		Different feature sets	Different feature sets with no LoG nor Wavelet features	Common feature sets
Local Control	SR radiomic	0.70 (0.65-0.72)	0.70 (0.65-0.72)	0.69 (0.65-0.72)
	PyR radiomic	0.63 (0.60-0.66)	0.62 (0.60-0.65)	0.63 (0.60-0.66)
	Clinical	0.57 (0.52-0.62)	0.57 (0.52-0.62)	0.57 (0.52-0.62)
	SR combined	0.62 (0.56-0.67)	0.62 (0.56-0.67)	0.63 (0.57-0.71)
	PyR combined	0.60 (0.54-0.65)	0.59 (0.52-0.61)	0.57 (0.52-0.61)
Distant Progression	SR radiomic	0.58 (0.56-0.60)	0.58 (0.56-0.60)	0.58 (0.57-0.61)
	PyR radiomic	0.55 (0.52-0.58)	0.57 (0.54-0.59)	0.57 (0.54-0.59)
	Clinical	0.51 (0.47-0.55)	0.51 (0.47-0.55)	0.51 (0.47-0.55)
	SR combined	0.56 (0.53-0.59)	0.56 (0.53-0.59)	0.56 (0.53-0.59)
	PyR combined	0.50 (0.46-0.54)	0.51 (0.48-0.56)	0.50 (0.47-0.56)
Overall Survival	SR radiomic	0.64 (0.61-0.67)	0.64 (0.61-0.67)	0.61 (0.57-0.64)
	PyR radiomic	0.63 (0.60-0.65)	0.54 (0.51-0.57)	0.54 (0.52-0.57)
	Clinical	0.59 (0.56-0.62)	0.59 (0.56-0.62)	0.59 (0.56-0.62)
	SR combined	0.65 (0.62-0-67)	0.65 (0.62-0-67)	0.62 (0.59-0.65)
	PyR combined	0.63 (0.60-0.66)	0.57 (0.53-0.60)	0.57 (0.54-0.60)

Table 4. **Summary of the models' results for each endpoint and feature-set scenario**. Q1=first-quartile or 25th percentile, Q3=third-quartile or 75th percentile, LoG=Laplacian of Gaussian, SR=SOPHiA Radiomics, PyR=PyRadiomics.

The corresponding notched boxplots for the 15 models are shown in Figure 2 a, b, and c.

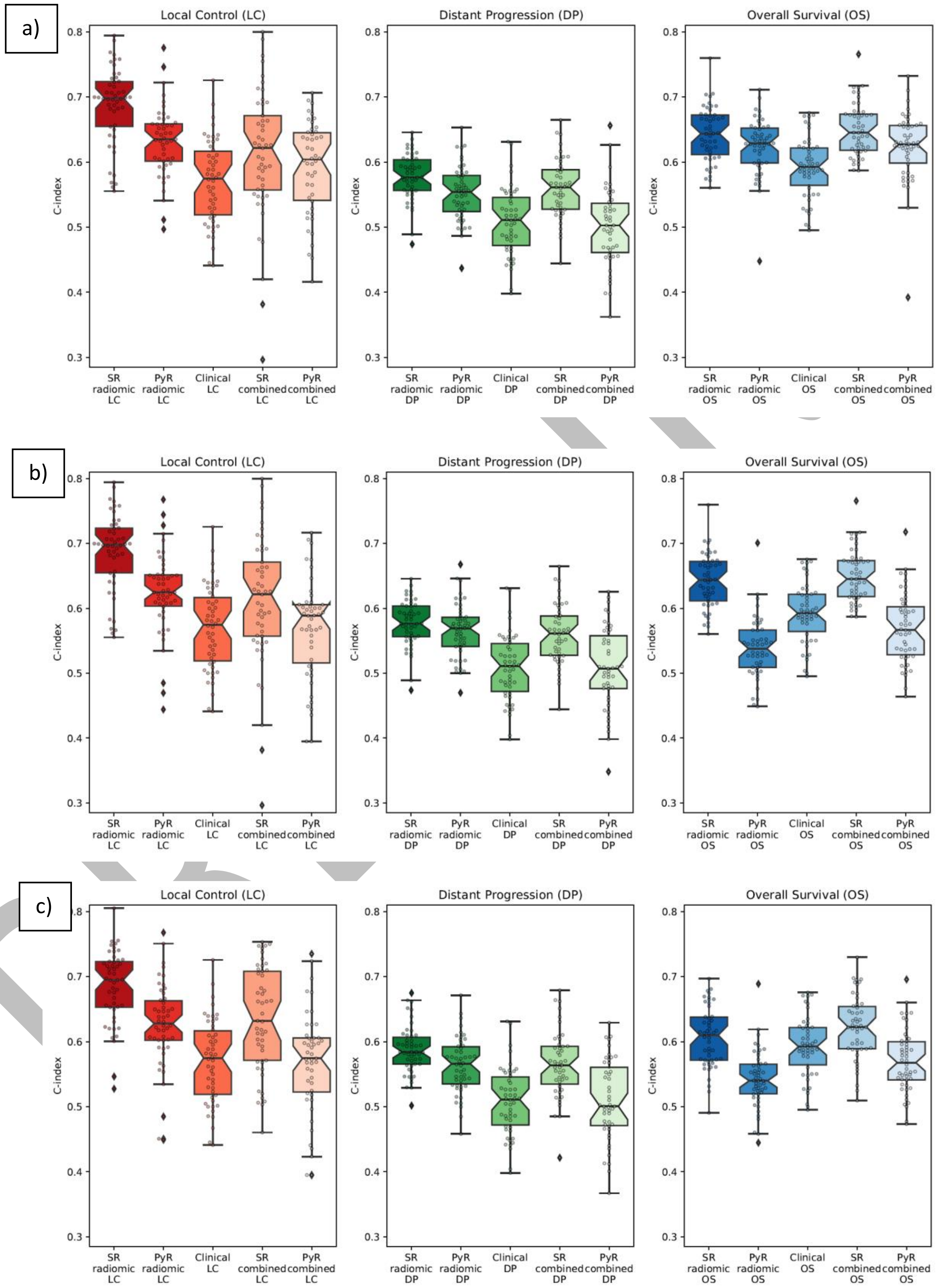

Figure 2. **Boxplots reporting the C-index performance of the 15 developed models using: (a) all the features extracted by the two platforms, (b) all the features excluding the LoG- and Wavelet-based features, (c) only the common features among the two platforms.** SR = SOPHiA Radiomics, PyR = PyRadiomics, C-index = concordance index.

The use of different platforms resulted in several mismatches, from three main viewpoints.

- **End-to-end performance**. In the case of different feature sets, the main gap was registered in radiomic models for LC prediction, where the SR model (median C-index=0.70) outperformed the PyR one (median C-index=0.63). This difference can be recognised as significant from the non-overlapping notches of the respective boxplots in Figure 2. That was true also for radiomic and combined models for DP. For the other models, inconsistencies were mild or not significant. From the side analyses, the exclusion of filter-based features did not mitigate the observed differences for LC prediction, attenuated the gap between radiomic models for DP, and led SR and PyR models to perform significantly different for OS. Training the models on the subset of common features enlarged the gap between LC combined models and between SR and PyR models for OS.
- **Statistical significance of RS**. From the radiomic models built with different feature sets, the two platforms largely disagreed on the statistical significance of RS (Table S.5). Both the exclusion of filter-based features and the usage of common features had almost no effect on the agreement of RS statistical significance for LC and DP, while increased the agreement for OS.
- **Significant clinical covariates in combined models**. The employment of different platforms led to a disagreement on the clinical covariates found significant in combined models developed upon the same clinical models (Table S.6). The major disagreement was recorded for OS models when using different feature sets, where no clinical feature in the SR combined model was deemed significant also in the PyR one, for the corresponding dataset split. Nevertheless, excluding filter-based features or using the common ones led to an alleviation of such differences, and that was true also for DP. Conversely, models for LC prediction exhibited poor concordance in all the three analyses.

4. Discussion

Assessment of performance fluctuations across different platforms poses serious implications for radiomic-based modelling. In fact, to spread trustworthiness among medical practitioners, these methods should behave accordingly when deployed at different institutions with possibly different extraction platforms.

In the present study, models integrating pre-RT T1-w MRI radiomic features and clinical data were developed for the prediction of LC, DP, and OS in patients treated with SRS for BMs from NSCLC. We assessed the variability in the performance of prognostic models when radiomic features are extracted from two platforms. By doing so, we compared three scenarios: (i) different feature sets are extracted from the platforms, simulating the case of different research hospitals carrying out independent analyses; (ii) similar to the former, but Log- and Wavelet- based features are excluded; and (iii) only the subset of features that are shared by the platforms are used.

Results of this study shed new light on how the extraction of radiomic features from the same images and segmentations by means of different platforms could, indeed, yield several discrepancies. The two platforms had different type and number of extracted features. Data revealed that such variations could be detrimental for C-index values of the final models, the most obvious example of which was the gap between C-indices across the two radiomic models for LC prediction. Furthermore, the main differences between the two platforms regarding the features' extraction parameters are reported in supplementary S7. The parameters used in SR, not known at the time of the principal analysis, became available upon our specific request to the developers of the software. Among the set of different parameters, those which might have influenced mostly the features extraction were resampling and normalization. For the others (such as -bin width, pre-crop and resegmentation range) a negligible impact on the numerical value of the features could be estimated. While the main aim of the present study was to evaluate the end to end performances of two different radiomic platforms, one open-source and customizable and the other one commercial and closed, the critical analysis of the impact of different parameters setting on features extraction and consequently on the final model should promote further studies to standardize the features extraction parameters allowing reproducible radiomic studies.

This work also disclosed that models built upon radiomic signatures coming from different platforms could strongly disagree on the statistical significance of RS. Furthermore, clinical features selected from clinical models could be deemed significant in the combined models to different extent of agreement based on whether the RS came from one platform or the other. Interestingly, the exclusion of filter-based features and/or the usage of common ones might not alleviate the observed differences, and several discrepancies were still found.

We proved that models' findings may change dramatically based on the platform used. It seems reasonable to assume that this may lead to acute consequences on drawing conclusions and comparing results across the literature.

Overall, across the 15 developed models in each of the three scenarios, the radiomic models based on SR features for LC resulted always the best performing ones (maximum median C-index = 0.70). Instead, DP analyses revealed that no specific model outperformed the others, and DP resulted the least predictable endpoint of the three for our dataset.

In addition, to the confirmation of poor robustness of predictive performance across different platforms [24], our study reveals a higher number of radiomic features does not necessarily imply better performance. In fact, in the scenario of different feature sets, the number of features extracted with PyR was almost six-fold w.r.t. SR. Knowledge in modelling and statistics suggests that a trade-off must be met in this case. Large, complex models composed of thousands of covariates are statistically more likely to discover unknown predictive factors, but at the same time are more prone to bias, overfitting, and computational demand. On the other hand, in compact and simple models deploying few hundreds of features, relevant factors may be missed, but computations are faster. This was true also in our case, where several consecutive clustering iterations were necessary to prune and deploy PyR features, whilst only one iteration was used for SR features.

Overall, our results are in line with previous works. Ji Zhang et al. [20] investigated the feasibility of a radiomics-based nomogram to predict OS from 195 NSCLC patients with BM treated with whole-brain RT. Features were extracted from pre-treatment CT images and selected with LASSO. By integrating radiomics and clinical features, authors reported a C-index of 0.66. Kothari et al. [34] claimed that radiomic models for OS prediction in NSCLC treated with curative RT have exhibited moderate prognostic capabilities so far (C-index random effects estimate was 0.57), and that standardised radiomics features should be considered.

Similar to our study, previous works attempted to conduct radiomic analyses about robustness and reliability of prognostic factors across different platforms. Foy et al.

[22] compared two in-house radiomics packages to two freely available packages using clinical images of various anatomic regions and imaging modalities and to determine sources of variations. The study demonstrated dramatic differences in computed radiomic features values across packages. These sources of variation included differences in image importation and pre-processing, algorithm implementation, as well as GLCM and feature-specific parameters. Fornacon-Wood et al. [24] investigated the effects of IBSI compliance, harmonization of calculation settings and platform version on the statistical reliability of radiomic features, and their corresponding ability to predict clinical outcomes. By leveraging three clinical datasets (108 head and neck cancer, 37 small-cell lung cancer, 47 NSCLC) they showed how the features identified as having significant relationship to survival varied between platforms, as did the direction of correlation. This issue was reported in our study too.

Our work remarked that radiomics practitioners should pay attention when drawing conclusions and provide a clear summary of software characteristics to foster results comparison across different centres. In addition, standardization initiatives to increase the generalizability and broaden the clinical applicability of radiomic models should be promoted.

This study presents some limitations. First, its retrospective nature, which opens the possibility of unforeseen variables and confounding biases, such as patient characteristics, imaging parameters, and treatment regimens for the primary NSCLC. Second, the small sample size, despite comparable to prior works, might have limited our ability to build more robust models. Third, a single feature reduction and selection methodology was followed, preventing possible comparison with other machine learning techniques available in the literature. Fourth, the source code for SR (commercial software) was not available, making it difficult to investigate the underlying mechanism and isolating the components such as pre-processing and algorithms; in addition, as SR is a closed commercial software, it is not possible to customise features extraction or to harmonise feature settings. Lastly, the selection of the worst responding lesion for the OS endpoint precludes any use of the model prior to treatment delivery. Both platforms claim to be compliant with the IBSI standard, but differences in their radiomics calculation approach may exist. Some claim that the default version of PyR is not completely compliant due to an issue on the implementation of the fixed bin size discretisation method [35]. Although to independently verify the IBSI compliance and to fix the related issues was outside of the scope of the present study, it may represent a suggestion for future work to assess whether differences are maintained.

The present study highlights how the choice of radiomic platform could impact the final performance of the models. This issue may be applicable to many imaging domains besides RT. Future paths of work would include the exploration of different methods for features reduction and classification, the harmonization of parameters across different platforms for features extraction, and the validation of our results with patients from external cohorts to evaluate generalizability.

References

[1] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. van Stiphout e P. Granton, «Radiomics: Extracting more information from medical images using advanced feature analysis.,» *Eur. J. Cancer,* vol. 48, pp. 441-446, 2012.

[2] L. Rinaldi, S. P. De Angelis, S. Raimondi, S. Rizzo, C. Fanciullo, C. Rampinelli, M. Mariani, A. Lascialfari, M. Cremonesi e R. Orecchia, «Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters,» *European Radiology Experimental,* vol. 6, n. 1, pp. 1-13, 2022.

[3] A. Midya, J. Chakraborty, M. Gönen, R. Do e A. Simpson, «Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility,» *J Med Imaging (Bellingham) 5,* pp. 11-20, 2018.

[4] R. Ger, J. Meier e R. Pahlka, «Effects of alterations in positron emission tomography imaging parameters on radiomics features,» *PLoS One,* vol. 14:e0221877, 2019.

[5] J. Wong, M. Baine e S. Wisnoskie, «Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer,» *Sci Rep,* vol. 11, n. 16328 , 2021.

[6] M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky e M. Friess, «Influence of inter-observer delineation variability on radiomics stability in different tumor sites,» *Acta Oncologica,* vol. 57, n. 8, pp. 1070-1074, 2018.

[7] M. Kocher, M. Ruge, N. Galldiks e P. Lohmann, «Applications of radiomics and machine learning for radiotherapy of malignant brain tumors,» *Strahlenther Onkol,* vol. 196, n. 10, pp. 856-867, 2020.

[8] R. Delgadillo, J. Ford, M. Abramowitz, A. Dal Pra, A. Pollack e R. Stoyanova, «The role of radiomics in prostate cancer radiotherapy,» *Strahlenther Onkol,* vol. 196, n. 10, pp. 900-912, 2020.

[9] V. Nardone, L. Boldrini, R. Grassi, D. Franceschini, I. Morelli, C. Becherini, M. Loi, D. Greto e I. Desideri, «Radiomics in the Setting of Neoadjuvant Radiotherapy: A New Approach for Tailored Treatment,» *Cancers (Basel),* vol. 13, n. 14, 2021.

[10] C. Giannitto, G. Marvaso, F. Botta, S. Raimondi, D. Alterio e D. Ciardo, «Association of quantitative MRI-based radiomic features with prognostic factors and recurrence rate in oropharyngeal squamous cell carcinoma,» *Neoplasma,* Vol. %1 di %267(6):1437-1446. doi: 10.4149/neo_2020_200310N249. Epub 2020 Aug 13. PMID: 32787435., 2020.

[11] S. Gugliandolo, M. Pepa, L. Isaksson, G. Marvaso, S. Raimondi, F. Botta, S. Gandini, D. Ciardo, S. Volpe, M. Cremonesi e M. Bellomi, «MRI-based radiomics signature for localized prostate cancer: a new clinical tool for cancer aggressiveness prediction? Sub-study of prospective phase II trial on ultra-hypofractionated radiotherapy (AIRC IG-13218),» *Eur Radiol.,* Vol. %1 di %231(2):716-728. doi: 10.1007/s00330-020-07105-z. Epub 2020 Aug 27. PMID: 32852590., 2021 Feb.

[12] L. Peng, V. Parekh, P. Huang, D. Lin, K. Sheikh e B. Baker, «Distinguishing true progression from radionecrosis after stereotactic radiation therapy for brain metastases with machine learning and radiomics,» *Int J Radiat Oncol Biol Phys,* vol. 102, p. 1236–1243, 2018.

[13] Z. Zhang, J. Yang, A. Ho, W. Jiang, J. Logan e X. Wang, « A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images,» *Eur Radiol.,* vol. 28, p. 2255–63, 2018.

[14] R. Ortiz-Ramon, A. Larroza, E. Arana e D. Moratal, «A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma.,» *Conf Proc IEEE Eng Med Biol Soc. ,* p. 493–6, 2017.

[15] H. Kniep, F. Madesta, T. Schneider, U. Hanning, M. Schonfeld e G. Schon, «Radiomics of brain MRI: utility in prediction of metastatic tumor type,» *Radiology. ,* vol. 290, p. 479–87, 2019.

[16] Z. Qian, Y. Li, Y. Wang, L. Li, R. Li e K. Wang, «Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers.,» *Cancer Lett,* vol. 451, p. 128–35, 2019.

[17] Y. Cha, W. Jang, M. Kim, H. Yoo, E. Paik e H. Jeong, «Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks,» *Anticancer Res. ,* vol. 38, p. 5437–45, 2018.

[18] M. Della Seta, F. Collettini, J. Chapiro, A. Angelidis, F. Engeling, B. Hamm e D. Kaul, «A 3D quantitative imaging biomarker in pre-treatment MRI predicts overall survival after stereotactic radiation therapy of patients with a singular brain metastasis,» *Acta Radiol,* vol. 60, n. 11, pp. 1496-1503, 2019.

[19] C. Huang, C. Lee e H. Yang, «Radiomics as prognostic factor in brain metastases treated with Gamma Knife radiosurgery,» *J Neurooncol,* vol. 146, p. 439–449.

[20] J. Zhang, «Computer Tomography Radiomics-Based Nomogram in the Survival Prediction for Brain Metastases From Non-Small Cell Lung Cancer Underwent Whole Brain Radiotherapy,» *Frontiers in Oncology,* 2021.

[21] A. Zwanenburg, M. Vallières, M. Abdalah e H. J. W. Aerts, «The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping,» *Radiology,* vol. 295, n. 2, pp. 328-338, 2020.

[22] J. R. Foy, H. KR Li, M. Giger, H. Al-Hallaq e S. Armato, «Variation in algorithm implementation across radiomics software,» *Journal of medical imaging (Bellingham, Wash.),* vol. 5, n. 4, 2018.

[23] M. Bogowicz, R. Leijenaar e S. Tanadini-Lang, «Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models,» *Radiotherapy and Oncology,* vol. 125, n. 3, pp. 385 - 391, 2017.

[24] I. Fornacon-Wood, H. Mistry e C. Ackermann, «Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform,» *Eur Radiol,* vol. 30, p. 6241–6250, 2020.

[25] Y. Chang, K. Lafata, C. Wang, X. Duan, R. Geng, Z. Yang e F. F. Yin, «Digital phantoms for characterizing inconsistencies among radiomics extraction toolboxes,» *Biomedical physics & engineering express,* vol. 6, n. 2, 2020.

[26] J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. Fillion-Robin, P. S. e H. J. Aerts, «Computational radiomics system to decode the radiographic phenotype,» *Cancer Res,* vol. 77, pp. 104-107, 2017.

[27] S. GENETICS, «SOPHiA Radiomics v2.1». Technopole Izarbel - 231 Allée Fauste d'Elhuyar, 64210 Bidart, France Brevetto www.sophiagenetics.com.

[28] E. L. Kaplan e P. Meier, «Nonparametric Estimation from Incomplete Observations,» *Journal of the American Statistical Association,* vol. 53, n. 282, pp. 457-481, 1958.

[29] F. Harrell, R. Califf, D. Pryor, K. Lee e R. Rosati, «Evaluating the yield of medical tests.,» *JAMA,* vol. 247, n. 18, pp. 2543-6, 1982 .

[30] J. H. Ward, «Hierarchical Grouping to Optimize an Objective Function,» *Journal of the American Statistical Association,* vol. 58, p. 236–244, 1963.

[31] R. Tibshirani, «Regression Shrinkage and Selection via the Lasso,» *Journal of the Royal Statistical Society.,* vol. 58, n. 1, pp. 267-288, 1996.

[32] D. R. Cox, «Regression models and life tables (with discussion).,» *Journal of the Royal Statistical Society,* vol. 34, n. B, pp. 187-220, 1972.

[33] A. Depeursinge, V. Andrearczyk, P. Whybra, J. van Griethuysen, H. Müller e R. Schaer, «Standardised convolutional filtering for radiomics,» *arXiv [csCV],* 2020.

[34] G. Kothari, J. Korte, E. J. Lehrer, N. G. Zaorsky, S. Lazarakis, T. Kron, N. Hardcastle e S. Siva, «A systematic review and meta-analysis of the prognostic value of radiomics based models in non-small cell lung cancer treated with curative radiotherapy,» *Radiotherapy and Oncology,* vol. 155, pp. 188-203, 2021.

[35] A. Bettinelli, F. Marturano, M. Avanzo, E. Loi, E. Menghi, E. Mezzenga, G. Pirrone, A. Sarnelli, L. Strigari, S. Strolin e M. Paiusco, «A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools,» *Radiology,* pp. 533-541, 2022.