External validation of an ^{18}F-FDG-PET radiomic model predicting survival after radiotherapy for oropharyngeal cancer

Martina Mori[1] · Chiara Deantoni[2] · Michela Olivieri[1] · Emiliano Spezi[1,2] · Anna Chiara[2] · Simone Baroni[2] · Maria Picchio[3,4] · Antonella Del Vecchio[1] · Nadia Gisella Di Muzio[2,6] · Claudio Fiorino[1] · Italo Dell'Oca[2]

Abstract

Purpose/objective The purpose of the study is to externally validate published ^{18}F-FDG-PET radiomic models for outcome prediction in patients with oropharyngeal cancer treated with chemoradiotherapy.

Material/methods Outcome data and pre-radiotherapy PET images of 100 oropharyngeal cancer patients (stage IV:78) treated with concomitant chemotherapy to 66–69 Gy/30 fr were available. Tumors were segmented using a previously validated semi-automatic method; 450 radiomic features (RF) were extracted according to IBSI (Image Biomarker Standardization Initiative) guidelines. Only one model for cancer-specific survival (CSS) prediction was suitable to be independently tested, according to our criteria. This model, in addition to HPV status, SUVmean and SUVmax, included two independent metafactors (F_i), resulting from combining selected RF clusters. In a subgroup of 66 patients with complete HPV information, the global risk score R was computed considering the original coefficients and was tested by Cox regression as predictive of CSS. Independently, only the radiomic risk score R_F derived from F_i was tested on the same subgroup to learn about the radiomics contribution to the model. The metabolic tumor volume (MTV) was also tested as a single predictor and its prediction performances were compared to the global and radiomic models. Finally, the validation of MTV and the radiomic score R_F were also tested on the entire dataset.

Results Regarding the analysis of the subgroup with HPV information, with a median follow-up of 41.6 months, seven patients died due to cancer. R was confirmed to be associated to CSS (p value = 0.05) with a C-index equal 0.75 (95% CI=0.62–0.85). The best cut-off value (equal to 0.15) showed high ability in patient stratification (p=0.01, HR=7.4, 95% CI=1.6–11.4). The 5-year CSS for R were 97% (95% CI: 93–100%) vs 74% (56–92%) for low- and high-risk groups, respectively. R_F and MTV alone were also significantly associated to CSS for the subgroup with an almost identical C-index. According to best cut-off value (R_F>0.12 and MTV>15.5cc), the 5-year CSS were 96% (95% CI: 89–100%) vs 65% (36–94%) and 97% (95% CI: 88–100%) *vs* 77% (58–93%) for RF and MTV, respectively. Results regarding RF and MTV were confirmed in the overall group.

Conclusion A previously published PET radiomic model for CSS prediction was independently validated. Performances of the model were similar to the ones of using only the MTV, without improvement of prediction accuracy.

Keywords Head neck cancer · Oropharyngeal cancer · PET · Predictive models · Radiomics · Imaging biomarkers

✉ Claudio Fiorino
fiorino.claudio@hsr.it

[1] School of Engineering, Cardiff University, Cardiff, UK

[2] Department of Medical Physics, Velindre Cancer Centre, Cardiff, UK

[3] Department of Nuclear Medicine, San Raffaele Scientific Institute, Milan, Italy

[4] Vita-Salute San Raffaele University, Milan, Italy

[1] Department of Medical Physics, San Raffaele Scientific Institute, Milano, Italy

[2] Department of Radiotherapy, San Raffaele Scientific Institute, Milano, Italy

Introduction

Head and neck cancer (HNC) is the sixth most common malignancy, with an incidence of about 650,000 cases and 330,000 deaths annually worldwide [1, 2]. It continues to be a clinically challenging problem because of several factors, including delayed detection, heterogeneous tumor subtypes that respond differently to treatment, difficult anatomical locations leading to adverse events, risk factors resulting in tumor recurrence, and comorbidities [3]. Although technologies are reaching their developmental peak, the question that still remains is inside the tumor that dictates different treatment responses from individuals despite similar tumor classification and/or clinical characteristics. Therefore, the interest in exploring biomarkers that could reliably and individually predict the tumor response to treatment is high. Quantitative extraction of high-dimensional data from medical images dealt with the field of Radiomics, a highly promising diagnostic, prognostic, and predictive tool for cancer characterization. Ideally, the integration of multiple-omics, i.e., "panomics" data (genomics, transcriptomics, proteomics, metabolomics, etc.), could efficiently unravel biological mechanisms [4–6]; in this context, radiomics may quantify tumor's texture, shape, and other geometric features, aiming to characterize tumor behavior, ideally functioning as a non-invasive, low-cost bridge between "biology" and "clinic" at individual level.

The translation of radiomic biomarkers into standard cancer care, to support treatment decision-making, involves the development of prediction models. Nowadays, there are several studies dealing with the development of models of outcome prediction for HNC [7–19] with the general impression that the scientific community is dragged by the current trend of a chaotic run in developing in-house models, mostly with limited validation. In addition, it is well known that still several aspects may impact the reliability of radiomic features (RF) as delineation, image acquisition/reconstruction, and bin size. Due to this lack of standardization, robustness studies are needed to assess the sensitivity of RF. A document with standardized feature definition was recently provided by the Image Biomarker Standardization Initiative (IBSI) [20] and is gradually becoming a reference guide. This implies that prognostic models based on RF analyses not IBSI compliant are expected to be discarded in the near future. In addition, models may intrinsically be more generalizable and easier to be adopted if the variables included are few, simple, and interpretable [21, 22].

To date, many studies were conducted in order to assess prognostic value of RF. Importantly, very few investigations considered functional imaging, primarily PET [7, 9, 19], that should in principle be more suitable in capturing tumor biological characteristics, potentially associated to a worse outcome.

CT radiomics was much more explored with several recent studies dealing with large cohorts [10–18], although external validation studies remain very rare and most investigations showed a clear correlation between RF-based scores and CT-based tumor volume [9–11, 13].

At our institute, PET imaging was introduced in planning optimization since mid-00s' [23, 24] and IBSI compliant procedures for radiomic analyses and outcome prediction studies in radiotherapy were implemented in the last years [25]. In this context, the aims of the current study were (1) to select from published PET-FDG radiomics prognostic models the ones, IBSI-consistent, considered to be suitable for an external validation on our population of patients treated for oropharyngeal cancer with radio-chemotherapy with PETFDG available; (2) to validate such models to possibly predict local recurrence (LRFS), distant metastasis (DRFS), and overall survival (CSS); and (3) to compare the performance of such models against the PET-based metabolic tumor volume (MTV) used as a single outcome predictor.

Materials and methods

Selection of published models

An extensive literature review was preliminary conducted to define suitable models for the current validation study. PubMed was used with the following search query: ("head neck"[Journal] OR ("head"[All Fields] AND "and"[All Fields] AND "neck"[All Fields]) OR "head and neck"[All Fields]) AND "pet"[All Fields] AND ("radiomic"[All Fields] OR "radiomics"[All Fields]) AND ("outcome"[All Fields] OR "outcomes"[All Fields]). Synonyms "cancer," "tumor," etc. were intentionally not used in the search to increase the comprehensiveness/inclusiveness of the search and to screen as much as possible the available literature on HNC. No date limit was used, and the search was updated until January 2022. As reported in the Supplementary

material (Table S1), 24 articles were selected. Papers were excluded because focused on hypoxia (5), reviews (1), works including the development of models based on imaging acquired during the treatment (4), presence of surgery (1), works based on (68)GA-DOTATOTE (1), RF extracted from CT and PET fusion (2), and no stated declaration of compliance with the IBSI guidelines (8). Finally, only one paper from the group of Martens [18] was selected: the resulting models were internally validated, in accordance with the TRIPOD level 2a of validation [26], while no independent validation is available.

Patient dataset

Outcome data and pre-radiotherapy PET images of 100 oropharyngeal cancer patients (stage IV: 78/100) treated at our institute between 2006 and 2021 according to an internal protocol delivering moderate hypo-fractionation (66 Gy/30 fr) were available: details of contouring, planning, and delivery procedures may be found elsewhere [23, 24]. Current study was approved by the Institutional Ethical Committee (n° 12/INT/2022). All patients underwent a planning 18F-fluorodeoxyglucose (FDG) positron emission tomography/computed tomography (PET/CT) at most one month before Radiotherapy, to assess MTV, and were treated with Helical TomoTherapy (HiArt 2, Accuray Inc.). Written informed consent for the execution of PET/CT and anonymous publication of disease-related information was signed by each patient. All patients were treated with SIB delivering 54 Gy (1.8 Gy/fr), 66 Gy (2.2 Gy/fr), and 69 Gy (2.3 Gy/fr) in 30 fractions on PTV-N, PTV-T, and MTV, respectively; in 84% of patients, concomitant CDDP chemotherapy (at least 200 mg/m^2 total dose) or cetuximab was also administered. For a subgroup of 66 patients, the HPV information was available. Table 1 summarizes the patient characteristics of both, the subgroup and of the complete dataset; patients were staged according to American Joint Committee on Cancer (AJCC) staging manual 7th edition.

Table 1 Patient's characteristics

	Subgroup of 66 patients with HPV data	Complete dataset of 100 patients
Age (years), (range)	65 (38–84)	65 (38–89)
Gender (male vs female)	51 vs 15	68 vs 32
Smoking history		
Yes	45 (68.1%)	60 (60.0%)
No	16 (24.4%)	26 (26.0%)
Missing	5 (7.5%)	14 (5.0%)
Alcohol history		
Yes	11 (16.7%)	23 (23.0%)
No	49 (74.2%)	61 (61.0%)
Missing	6 (9.1%)	16 (16.0%)
HPV status		
Positive	50 (75.8%)	50 (50.0%)
Negative	16 (24.2%)	16 (16.0%)
Missing	0 (0%)	34 (34.0%)

In the Martens et al. study, treatment consisted of a chemoradiotherapy (CRT) during a period of 7 weeks followed by 70 Gy in 35 fractions with concomitant cisplatin (100 mg/ m^2 on days 1, 22, and 43 of radiotherapy) or cetuximab (400 mg/m^2 loading dose followed by seven weekly infusions of 250 mg/m^2). Like in the Martens et al. study, loco-regional recurrence was measured from the end of CRT to the date of local or regional proven relapse. Metastases were defined as a distant location from the loco-regional primary tumor and lymph nodes. CSS time was measured from the end of CRT until death or the last follow-up date.

Image acquisition, target segmentation, and RF extraction

The characteristics of scanners and acquisition protocols as well as the differences with the Martens et al. study are reported in detail in the Supplementary material. Segmentation of tumor MTV was performed using the semi-automatic contour method, named "PET_Edge," based on a gradient edge search (MIM Software Inc., Cleveland, OH, USA). The method was previously tested as reproducible and accurate compared to manual segmentation [27]. In the Martens' study, delineation of primary tumors was performed semiautomatically on 18F-FDG-PET/CT using a 50% isocontour of the SUV-peak of the tumor volume. SUV was normalized to body weight. Since this part of the study was the

Clinical stage		
II	1 (1.5%)	3 (3.0%)
III	13 (19.7%)	19 (19.0%)
IV	52 (78.8%)	78 (78.0%)
Concomitant chemotherapy		
No	9 (13.6%)	16 (16.0%)
Cetuximab	9 (13.6%)	10 (10.0%)
Cisplatin	48 (72.8%)	74 (74.0%)

Zone Matrix 3D, Neighbors Grey Tone Difference Matrix 3D (NGTDM3D), and Grey Level Distance Zone Matrix 3D (GLDZM3D). In Fig. 1, a workflow of the whole pipeline was summarized.

Validating the Martens model and comparing performances against MTV

Martens et al. condensed the predictive RF in 8 independent meta-factors (F_i), consisting of a combination of selected RF with variable importance weight. According to their publications [18], F_i were built using the weights reported for each RF. A global score risk R was computed for DRFS (not including F_i), LRFS (including HPV, SUVmean, SUVpeak, F3, F4, F6) and CSS (including HPV, SUVmean, SUVmax, F1 and F5). Due to the limited availability of the only which could be done prospectively, all images were processed to reach conditions similar to those reported by Martens et al. Given the different activities administered to patients, we limited the adaptation to the process of all the PET images to the same voxel size of 4×4×4 mm^3. Images were discretized to a fixed bin number of 64, due to the improved reproducibility as reported by Tixier et al. [28] and confirmed in an ad hoc phantom study [29]. As Martens et al. used a fixed bin size (0.25 SUV) approach, potentially different results could derive from this different technical process. Similar to Martens et al., RF directly computed from the DICOM images were scaled to the interval [0, 1] to avoid a situation where the features with the largest scale dominate the analysis. DICOM files were imported to MATLAB using the Computational Environment for Radiological Research (https://cerr .github .io /CERR /) . RF extraction was performed with SPAARC Pipeline for Automated Analysis and Radiomics Computing (SPAARC [30, 31]) developed at Cardiff University School of Engineering. SPAARC complies with the IBSI guidelines [20]. We extracted 450 RF belonging to all the families included in IBSI: Morphology, Statistical, Intensity Histogram, Grey Level Co-occurrence Matrix 3D_average (GLCM3D_avg), Grey Level Co-occurrence Matrix 3D_combined (GLCM3D_comb), Grey Level Run Length 3D_average (GLRL3D_avg), Grey Level Run Length 3D_combined (GLRL3D_comb), Grey Level Size

HPV-status, we limited the validation of the Martens models to the subgroup of 66 patients with HPV information available, testing the global score risk R prediction by Cox regression. R was computed as a liner combination of the original coefficients of the Martens study and the covariates selected. Moreover, a radiomic risk score R_F involving only F_i was tested on both the subgroup and on the complete dataset. It was computed as well using the original coefficients of the Martens model applied to F_i in order to compare results between the global and an “only radiomic” model. The resulting R and R_F indexes were then used to stratify risk according to the best cut-off value derived from the ROC analysis [25]. Kaplan-Meier test was finally performed. Due to the evidence that RF may be a surrogate of the tumor volume as reported for CT-based volumes [9–11, 13], the semi-automatically segmented MTV was

independently tested as a single potential predictor of outcome. The same procedures followed for the considered R and R_F scores to test their prediction performances were applied. The performances of the two approaches, Martens risk model (global and radiomic) vs MTV, were compared in terms of concordance index (C-index), hazard ratios (HR), and *p* value.

Results

The median follow-up was 42.7 months (IQR: 21–71). At the time of analysis, 69/100 patients were alive and 4, 13, and 15 events for LRFS, DRFS, and CSS were registered. Regarding the subgroup with HPV data information, the median follow-up was 41.6 months (IQR: 19.1–67.4). At the time of analysis, 41/66 patients were alive and 2, 10, and 7 events for LRFS, DRFS, and CSS were registered. In Martens et al., RF selected as predictive were found only for LRFS and CSS prediction. Due to the too small number of LRFS events in our population, only the model for CSS could be tested. The significant radiomic predictors for CSS in Martens' study were F1 and F5 [18]. F1 and F5 together with HPV, SUVmean, and SUVmax were included in the global model. The global risk score R was calculated by using the original coefficients derived by Martens on the subgroup with HPV data. The radiomic risk score R_F was calculated

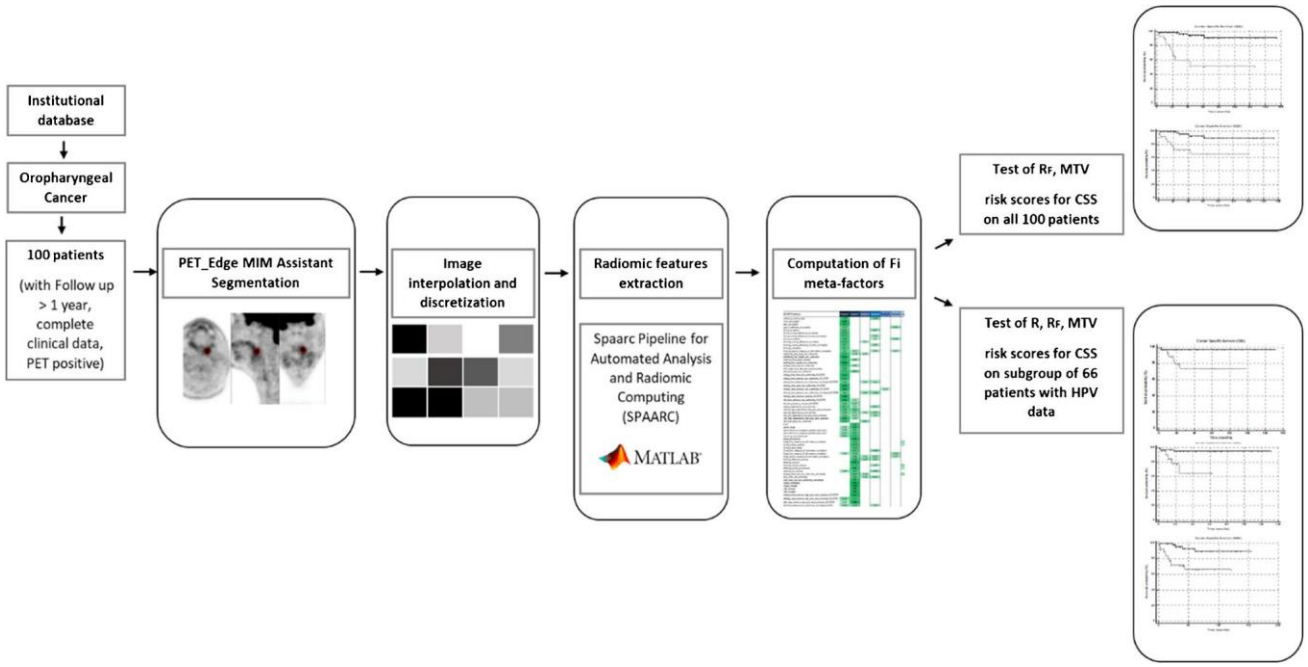

Fig. 1 Summary of the workflow followed for model's validation **Table 2** Cox regression of global, radiomic risk scores (R and R_F) and the MTV for the prediction of CSS on the same subgroup of patients with HPV information

Risk score	*b*	*p*	Exp(b)	C-index
Global risk score R	3.4962	0.050	12.99 (95% CI: 0.90–47.25)	0.75 (95% CI: 0.62–0.85)
Radiomic risk score R_F	3.6132	0.030	17.09 (95% CI: 1.41–58.44)	0.75 (95% CI: 0.62–0.85)
MTV	3.3524	0.041	28.57 (95% CI: 1.90–49.35)	0.76 (95% CI: 0.67–0.89)

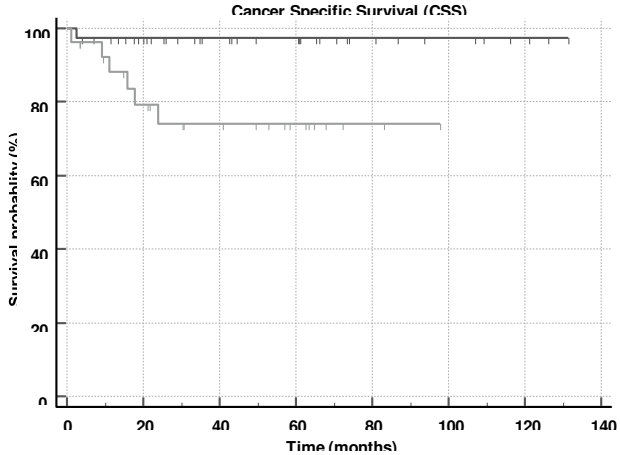

Fig. 2 Kaplan Meier curves of CSS patients' stratification according to the best cut-off for the risk score R (low risk: R<0.15 grey line; high risk: R>0.15 black line)

from F1 and F5 as well, but, as the coefficients referred to the "combined" model, we assumed that the relative weight of the two meta-factors was independent on the other variables. Then, we used the original coefficients renormalized to the value referred to F1 (1 and 1.04 for F1 and F5, respectively). From the Cox regression analysis, the resulting risk scores R and R_F resulted associated to CSS (Table 2) (p=0.05 and p=0.03, respectively) and for both the C-index was found equal to 0.75 (95% CI=0.62–0.85). According to the Youden criterion, the best cut-off for R and R_F were found >0.15 and >0.12 respectively, showing in both cases a high ability in patients' stratification, as depicted by the Kaplan Maier curves reported in Fig. 2 (p=0.01, HR=7.4, 95% CI=1.6–11.4 for R) and in Fig. 3 (p=0.006, HR=11.1, 95% CI=2.02–15.3 for R_F). The 5-year CSS for R were 97% (95% CI: 93–100%) vs 74% (56–92%) and for R_F 96% (95% CI: 89–100%) vs 65% (36–94%) for low- and high-risk groups, respectively.

MTV alone was significantly associated to CSS in this subgroup (p=0.04) with a C-index=0.76 (95% CI=0.67–0.89). When considering the best cut-off value (volume<15.5cc), patients were well stratified (p<0.0097, HR=7.5, 95% CI=1.6–34.4), and the 5-year CSS were 97% (95% CI: 88–100%) vs 77% (58–93%) for low- and high-risk groups, respectively (Fig. 4).

When looking to the complete dataset, the validation of the radiomic score R_F and the MTV was confirmed.

Fig. 3 Kaplan Meier curves of CSS patients' stratification according to the best cut-off for the radiomic risk score (low risk: R_F<0.12 grey line; high risk: R_F>0.12 black line)

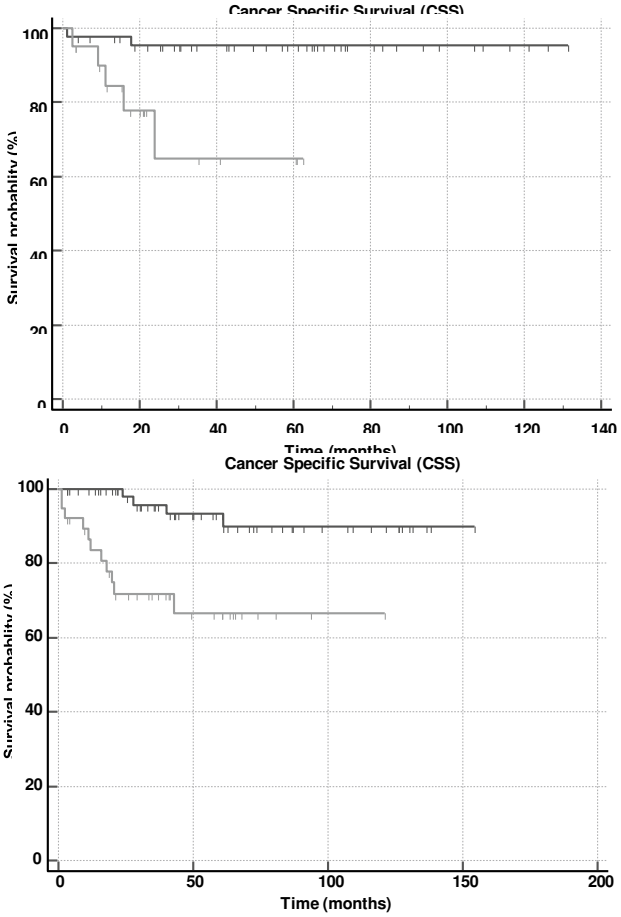

Fig. 4 Kaplan Meier curves of CSS patients' stratification according to the best cut-off of the semi-automatically segmented PET-based volume (low risk: MTV<15.5 cc grey line; high risk: MTV>15.5 cc black line)

The results are similar (C-index=0.75 (95% CI=0.66–0.83, p=0.008 for RF and C-index=0.76 (95% CI=0.67–0.85, p=0.008 for MTV). Results are summarized in the Supplementary material.

Discussion

Currently, pre-treatment imaging of head and neck cancers serves the purpose of evaluating primary tumor dimensions, anatomical extent, involvement of regional lymph nodes, and detecting distant metastases, which constitute the basis for staging and therapeutic choice. While PET/CT represents a mainstay of disease work-up, human visual interpretation cannot seize the full prognostic utility encoded in metabolic and structural bioimaging patterns. By capturing such bioimaging features, radiomic biomarkers may, in principle, improve stratification of patient risk groups and patient selection in better guiding personalized therapy. Quantitative imaging biodata reflecting tissue

density, texture patterns, lesion shape, and metabolic activity of primary tumors and metastatic cervical nodes may encode valuable information pertaining to tumor behavior with potential prognostic relevance.

To date, many studies were conducted to assess prognostic value of RF. Although functional imaging should in principle be more suitable in capturing tumor biology, CT radiomics was much more explored with several recent studies dealing with large cohorts [10–18].

Focusing on CT radiomic investigations with extensive validation, results are in part contradictory with few negative studies reporting no advantage in including radiomic features compared to clinical and geometrical/anatomical features into models [9, 10]. These results were explained by the strong correlation of most selected features with tumor volume that was able alone to perform predictions similarly to more complex radiomic scores, as previously shown by Welch et al. [11] in re-analyzing the early study by Aerts et al. [8]. The issue of the dependence of many RF on tumor volume is relevant and, due to this, results showing high performances of radiomic scores that did not consider this issue should be regarded with caution. Our study, despite the use of PET in place of CT, indirectly confirmed that tumor volume is likely to be the major predictor, making difficult to demonstrate an additional value of more complex radiomic patterns.

Among the positive studies using radiomic CT, the ones with the strongest ability of radiomic scores to stratify patients according to their prognosis were driven to identify few, highly predictive, features taken as the major predictors, once redundancy filters were applied. For instance, Elhalawani et al. [30] showed that the addition of a radiomic score composed of the combination of only two features ("intensity direct local range max" and "neighbor intensity difference 2.5 complexity") improved the prediction of local recurrences in a population of 465 oropharyngeal patients. Cozzi et al. [12] identified the combination of two-three features in stratifying 110 patients in high- and low-risk groups. Meneghetti et al. [13] recently showed that the combination of tumor volume with two independent radiomic features improved prediction of loco-regional relapses in a large population merging 6 German cohorts. They showed that, once properly managed, additional contribution of radiomic features not depending on volume can be detected. Similarly, Zhai et al. [14] showed that the combination of one feature related to the volume (least axis length) and one independent (gray level co-occurrence base correlation) extracted from positive nodes can carefully predict individual lymphnode failure in a group of 112 patients (with 558 analyzed nodes).

More sophisticated advanced machine learning and deep learning approaches to build radiomic scores have also been explored, most of them still using CT [15–17]. Among them, the most relevant is probably the one by Giraud et al. [17], due to their effort to make the resulting models for locoregional and overall survival interpretable though a graphic representation of the weight of many features (radiomic and clinical) included in the models. On the other hand, the large number of features compared to the number of patients could have generated some overfit. The same study still reported the "shape voxel volume" features, strongly correlated with tumor volume, as the most prominent predictor. CT radiomic in the field of HN cancer was also investigated in HPV classification [33] or in assessing outcome of therapies other than radiotherapy [34, 35].

When considering PET imaging, very few investigations were reported: for instance, Ger et al. [9] in the previously discussed negative study tested also PET-FDG-related features, and none was retained in the final model. Feliciani et al. [19] found one single feature ("low-intensity longrun emphasis") able to predict outcome in a heterogeneous cohort of 129 patients. In the pioneer work by Vallieres et al. [7], several PET-FDG features were combined to develop radiomic models in predicting outcome. The combination of features extracted by using different processing parameters made these models hard to apply. In addition, PET features did not demonstrate any additional benefit compared to clinical/volume variables neither to CT features.

More in general, the possibility to replicate performances is a critical issue for radiomic models. As a matter of fact, our exercise was in the direction of selecting models that could be replicated, even considering the consistency with IBSI guidelines. Not by chance, it is hard to find models that explicitly satisfy these criteria, also due to the quite recent publications of these guidelines that appeared only in 2020 [20]. After a careful selection, the paper by Martens et al. [18] was found to satisfy them and chosen for independent validation on our institutional cohort. Their study applied a quite innovative, cluster-based, analysis that allowed to identify different scores representing "meta-features" independently predictive of outcome. In particular, due to our available data, we focused on the "radiomic-only" model for overall survival. Interestingly, the above-mentioned approach made possible to split the contributions of features depending on tumor volume (as identified by a previously validated semi-independent method based on SUV gradient, F1) against the ones not depending on volume (F5).

Our results show a good replication of the previously reported ability of the Martens risk score R in stratifying patients in low and high risk based on cancer-specific survival.

On the other hand, very importantly, R showed performances similar to the ones of the radiomic-only score R_F, suggesting that the additional benefit of HPV and other variables SUV-related could be already included in the radiomics information, according to a recent published work [36]. On the other hand, the limited statistics cannot permit to fully clarify this issue.

If considering MTV, the performances are quite similar in terms of C-index; when assessing best cut-off values, MTV showed a similar trend in stratifying patients compared to R and R_F. Very importantly, results regarding R_F and MTV were confirmed on the complete dataset, corroborating our positive results.

Of note, the outcome prediction power of MTV was already reported in other investigations [37–40] and confirms the potential of using a simple, reproducible, operatorindependent parameter to classify patients according to their outcome. Of note, MTV can be reliably obtained semi-automatically [27], as done in current study, overcoming the issue of contouring uncertainty, so relevant in the case of CT.

To the best of our knowledge, this is the first study reporting an independent validation of a published PET-based radiomic model predicting outcome in patients treated with radio-chemotherapy for head-neck cancer. In our population, the performances of such radiomic score in predicting CSS were not significantly superior to using just the MTV. The unavailability of the HPV status for all patients limited the possibility to replicate the prediction of the combined model incorporating this and other factors.

Conclusions

Pre-treatment PET/CT radiomics biomarkers may provide complementary prognostic value for oropharyngeal cancer via systematic quantification of tissue density, texture patterns, lesion geometry, and metabolic properties. We independently confirmed the value of a previously published model based on clinical data and radiomic meta-factors for CSS prognostication and risk stratification. The reproducibility of the dataset used, as results, probably depict the prognostic potentials of radiomic biomarkers for CSS in a realistic fashion. However, a similar predictive power was reached if using only the (semi-automatically segmented) MTV, suggesting that the additional benefit of more complex PET RF-based signatures remains to be demonstrated and, consistently with recent CT-based radiomic results, could be limited. Despite this promising result, more studies are needed to evaluate the predictive power of different PET RF-based signatures and their potential benefit to clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi .org /10 .1007 /s00259 -022 -06098- 9.

Author contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by M Mori, C Deantoni, M Olivieri, A Chiara, S Baroni, C Fiorino, and I Dell'Oca. The first draft of the manuscript was written by M Mori and C Fiorino, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The project was supported by an AIRC (Associazione Italiana per la Ricerca sul Cancro) grant (IG23150).

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of San Raffaele Scientific Institute (March 10th, 2022/N° 12/INT/2022).

Consent to participate/to publish Written informed consent for the execution of PET/CT and anonymous publication of disease-related information was signed by each patient.

Competing interests The authors declare no competing interests.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjoemataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin. 2017;67:7–30.
3. Marcu LG, Boyd C, Bezak E. Feeding the data monster: data science in head and neck cancer for personalized therapy. J Am Coll Radiol. 2019;16:12.
4. El Naqa I. Biomedical informatics and panomics for evidencebased radiation therapy. WIREs Data Mining Knowl Discov. 2014;4:327–40.
5. Ebrahim A, Brunk E, Tan J, et al. Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun. 2016;7:13091.
6. Vallières M, Zwanenburg A, Badic B, et al. Responsible radiomics research for faster clinical translation. Nucl Med. 2018;59(2):189–93. https://doi .org /10 .2967 /jnumed .117 .20050 1.
7. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117. https://doi .org /10 .1038 /s41598 -017 -10371- 5.
8. Aerts HJWL, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.

9. Ger R, Zhou S, Elgohari B, et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- And PET-imaged head and neck cancer patients. PLoS ONE. 2019;14(9):e0222509.
10. Keek S, Sanduleanu S, Wesseling F, et al. Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemoradiotherapy. PLoS ONE. 2020;15(5):e0232639.
11. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol. 2019;130:2–9.
12. Cozzi L, Franzese C, Fogliata A, et al. Predicting survival and local control after radiochemotherapy in locally advanced head and neck cancer by means of computed tomography based radiomics. Strahlenther Onkol. 2019;195(9):805–18.
13. Meneghetti AR, Zwanenburg A, Leger A, et al. Definition and validation of a radiomics signature for loco-regional tumour control in patients with locally advanced head and neck squamous cell carcinoma. Clin Transl Radiat Oncol. 2021;26:62–70.
14. Zhai TT, Langendijk JA, van Dijk LV, et al. Pre-treatment radiomic features predict individual lymph node failure for head and neck cancer patients. Radiother Oncol. 2020;146:58–65.
15. Volpe S, Pepa M, Zaffaroni M, et al. Machine learning for head and neck cancer: a safe bet? A clinically oriented systematic review for the radiation oncologist. Front Oncol. 2021;11:772663.
16. Le WT, Vorontsov E, Romero FP, et al. Cross-institutional outcome prediction for head and neck cancer patients using selfattention neural networks. Sci Rep. 2022;12(1):3183.
17. Giraud P, Giraud P, Gasnier A, et al. Radiomics and machine learning for radiotherapy in head and neck cancers. Front Oncol. 2019;9:174.
18. Martens RM, Koopman T, Noij DP, et al. Predictive value of quantitative ^{18}F-FDG-PET radiomics analysis in patients with head and neck squamous cell carcinoma. EJNMMI Res. 2020;10:102.
19. Feliciani G, Fioroni F, Grassi E et al. Radiomic profiling of head and neck cancer: 18F-FDG PET texture analysis as predictor of patient survival. Contrast Media Mol. Imaging, 2018 3574310
20. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based Phenotyping. Radiology. 2020;295(2):328–38. https://doi .org /10 .1148 /radiol . 202019114 5.
21. Kundu S. AI in medicine must be explainable. Nat Med. 2021;27:1328.
22. Fiorino C, Rancati T. Artificial intelligence applied to medicine: there is an “elephant in the room.” Phys Med. 2022;98:8–10.
23. Fiorino C, Dell’Oca I, Pierelli A, et al. Simultaneous integrated boost (SIB) for nasopharynx cancer with helical tomotherapy: a planning study. Strahlenther Onkol. 2007;39:497–505.
24. Widesott L, Pierelli A, Fiorino C, et al. Intensity-modulated proton therapy versus helical tomotherapy in nasopharynx cancer: planning comparison and NTCP evaluation. Int J Radiat Oncol Biol Phys. 2008;72:589–96.
25. Mori M, Passoni P, Incerti E, et al. Training and validation of a robust PET radiomic-based index to predict distant-relapsefree-survival after radio-chemotherapy for locally advanced pancreatic cancer. Radiother Oncol. 2020;153:258–64.
26. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual Prognosis or diagnosis (TriPod): the TriPod statement. Ann intern Med. 2015;162:55–63.
27. Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell’Oca I, Fallanca F, Passoni P, Vanoli EM, Calandrino R, Di Muzio N, Picchio M, Fiorino C. Quantifying the robustness of [18F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Med. 2018;49:105–11.
28. Tixier F, Hatt M, Le Rest CC, et al. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med. 2012;53(5):693–700.
29. Presotto L, Bettinardi V, De Bernardi E, et al. PET textural features stability and pattern discrimination power for radiomics analysis: an “ad-hoc” phantoms study. Phys Med. 2018;50:66–74.
30. Whybra P, Parkinson C, Foley K, et al. Assessing radiomic feature robustness to interpolation in 18 F-FDG PET imaging. Scientific Reports. 2019;9(1):9649.
31. Piazzese C, Foley K, Whybra P, et al. Discovery of stable and prognostic CT-based radiomic features independent of contrast administration and dimensionality in oesophageal cancer. PLOS ONE. 2019;14(11):e0225550. https://doi .org /10 .1371 /jour nal .pone .0225 55 0.
32. Elhalawani H, Kanwar A, Mohamed ASR, et al. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. Scientific Reports. 2018;8(1):1524.
33. Ou D, Blanchard P, Rosellini S, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. Oral Oncology. 2017;71:150–5.
34. Sun R, Limkin EJ, Vakalopoulou M, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. The Lancet Oncology. 2018;19(9):1180–91.
35. Masson I, Da-ano R, Lucia F, et al. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal cancers. Med Phys. 2021;48(7):4099–109.
36. Lv Wenbing, Hui Xu, Han Xu, et al. Context-aware saliency guided radiomics: application to prediction of outcome and HPV-status from multi-center PET/CT images of head and neck cancer. Cancers. 2022;14:1674.
37. Picchio M, Kirienko M, Mapelli L, et al. Predictive value of F18-FDGPET/CT for the outcome of F18-FDG PET-guided radiotherapy in patients with head-neck cancer. Eur J Nucl Med Mol Imaging. 2014;41:21–31.
38. Schwartz DL, Harris J, Yao M, et al. Metabolic tumor volume as a prognostic imaging-based biomarker for head-and-neck cancer: pilot results from Radiation Therapy Oncology Group protocol 0522. Int J Radiat Oncol Biol Phys. 2015;91:721–9.
39. Rijo-Cedeño J, Mucientes J, Álvarez O, et al. Metabolic tumor volume and total lesion glycolysis as prognostic factors in head and neck cancer: systematic review and meta-analysis. Head and Neck. 2020;42:3744–54.
40. Won Kim J, Oh JS, Roh JL, Kim JS, Choi SH, Nam SY, Sang Kim SY. Prognostic significance of standardized uptake value and metabolic tumour volume on 18F-FDG PET/CT in oropharyngeal squamous cell carcinoma. European J Nuclear Medicine and Molecular Imaging. 2015;42:1353–61.