Comparison of computed tomography image features extracted by radiomics, self-supervised learning and end-to-end deep learning for outcome prediction of oropharyngeal cancer

Ma, Baoqiang; Guo, Jiapan; Chu, Hung; van Dijk, Lisanne V.; van Ooijen, Peter M.A.; Langendijk, Johannes A.; Both, Stefan; Sijtsema, Nanna M.

Published in:
Physics and imaging in radiation oncology

DOI:
10.1016/j.phro.2023.100502

Publication date:
2023

Physics and Imaging in Radiation Oncology 28 (2023) 100502

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Comparison of computed tomography image features extracted by radiomics, self-supervised learning and end-to-end deep learning for outcome prediction of oropharyngeal cancer

Baoqiang Ma [a,*], Jiapan Guo [a,b,c], Hung Chu [a,b,d], Lisanne V. van Dijk [a,e], Peter M.A. van Ooijen [a,b], Johannes A. Langendijk [a], Stefan Both [a], Nanna M. Sijtsema [a]

[a] Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[b] Machine Learning Lab, Data Science Center in Health (DASH), Groningen, Netherlands
[c] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence , University of Groningen, Groningen, Netherlands
[d] Center for Information Technology, University of Groningen ,Groningen, Netherlands
[e] Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX USA

ARTICLE INFO

Keywords:
Oropharynx carcinoma
Radiomics
Deep learning
Self-supervised learning
Prognostic modeling

ABSTRACT

Background and purpose: To compare the prediction performance of image features of computed tomography (CT) images extracted by radiomics, self-supervised learning and end-to-end deep learning for local control (LC), regional control (RC), locoregional control (LRC), distant metastasis-free survival (DMFS), tumor-specific survival (TSS), overall survival (OS) and disease-free survival (DFS) of oropharyngeal squamous cell carcinoma (OPSCC) patients after (chemo)radiotherapy.

Methods and materials: The OPC-Radiomics dataset was used for model development and independent internal testing and the UMCG-OPC set for external testing. Image features were extracted from the Gross Tumor Volume contours of the primary tumor (GTVt) regions in CT scans when using radiomics or a self-supervised learning-based method (autoencoder). Clinical and combined (radiomics, autoencoder or end-to-end) models were built using multivariable Cox proportional-hazard analysis with clinical features only and both clinical and image features for LC, RC, LRC, DMFS, TSS, OS and DFS prediction, respectively.

Results: In the internal test set, combined autoencoder models performed better than clinical models and combined radiomics models for LC, RC, LRC, DMFS, TSS and DFS prediction (largest improvements in C-index: 0.91 vs. 0.76 in RC and 0.74 vs. 0.60 in DMFS). In the external test set, combined radiomics models performed better than clinical and combined autoencoder models for all endpoints (largest improvements in LC, 0.82 vs. 0.71). Furthermore, combined models performed better in risk stratification than clinical models and showed good calibration for most endpoints.

Conclusions: Image features extracted using self-supervised learning showed best internal prediction performance while radiomics features have better external generalizability.

1. Introduction

Head and neck cancer (HNC) is a common cancer type worldwide and is commonly treated with surgery, (chemo)radiotherapy, or both. Oropharyngeal squamous cell carcinoma (OPSCC), an important type of HNC, includes human papillomavirus (HPV) related (HPV +) and HPV-unrelated (HPV-) tumors, which are mostly related to cigarette and alcohol use. 5-year overall survival (OS) rate in patients with HPV + tumors is generally better (75 %-80 %) than in those with HPV-negative tumors (45 %-50 %) [1]. To allow for more individualized treatment choices in the future, it is necessary that outcome prediction models with good performance become available for different endpoints like local control (LC), regional control (RC), distant metastasis free survival (DMFS) and OS.

Clinical parameters such as HPV-status, age, gender, T-stage, N-stage and smoking status have been identified as prognostic factors of OS [2–12], progression-free survival (PFS) [7,8,11] and locoregional control (LRC) [3] for OPSCC patients. The risk stratification system for OS in

* Corresponding author at: Department of Radiation Oncology, University Medical Center Groningen, PO Box 30001, 9700RB Groningen, The Netherlands.
E-mail address: b.ma@umcg.nl (B. Ma).

https://doi.org/10.1016/j.phro.2023.100502
Received 3 August 2023; Received in revised form 2 October 2023; Accepted 17 October 2023
Available online 7 November 2023

OPSCC patients based on HPV-status, pack years of smoking, tumor and nodal stages proposed by Ang et al. [2] has a good performance and is commonly used clinically. However, the availability and performance of clinical models for other endpoints like local and regional control is still limited [3,13]. Some studies showed that the addition of radiomic features extracted from MRI [3] or PET/CT [14–17] improved the performance of prediction models for LRC and OS compared to that of clinical models. However, these hand-crafted radiomics features are limited describing the inherent characteristics of tumors.

Compared to radiomics, convolutional neural networks (CNNs) can extract more representative and descriptive image features and have been successfully applied in various tasks such as image synthesis [18–20], super-resolution [21,22] and segmentation [23,24]. Recently, researchers have applied CNNs in radiotherapy applications including automatic segmentation [25,26], treatment planning [27] and outcome prediction [28–30]. For the outcome prediction of OPSCC, Fujima et. al. used CNNs to extract PET image features to predict local treatment outcomes [31]. Moreover, Cheng et. al. proposed a fully automatic tumor segmentation and OS prediction tool of OPSCC based on CNNs [32]. Naser et. al. applied a DenseNet based method to extract features from CT, PET, (gross tumor volumes of the primary tumor) GTVt and clinical data together to predict PFS of OPSCC [33]. Our previous studies extracted features from CT, PET and GTVt using self-supervised or end-to-end learning based methods and built deep learning models based on clinical and image features together for outcome prediction [34–36]. However, these works relied on PET images which are less available than planning CTs.

Based on CT only, Diamant et. al. built 2D CNN models with an input of 2D central tumor slice of pretreatment CTs [37], and achieved better prediction than previous radiomics models [38] in the prediction of locoregional failure, distant metastasis (DM) and OS of HNC. Lombardo et. al extended this 2D CNN to 3D for time-to-event DM prediction [39]. Their 3D CNNs generally obtained good C-index values of around 0.80 in the validation sets and two of three independent test sets. Wang et. al proposed a 3D residual block-based model and achieved C-indexes of 0.77 and 0.64 for DM and OS prediction when inputting the 3D CT volume of the GTV (gross tumor volume) region [40]. The above studies included a comparison of radiomics models and end-to-end deep learning models. However, they investigated a limited set of endpoints and did not always compare the radiomics and CNN models to prediction models using clinical features only. In this study, we aimed to extract tumor image features from the pretreatment CT and investigated whether these image features can improve the performance of clinical parameters-based outcome prediction models for OPSCC patients. The predicted abilities of CT image features extracted by radiomics, self-supervised learning (autoencoder) and end-to-end deep learning, respectively, were investigated and compared for the prediction of local control (LC), regional control (RC), LRC, distant metastasis-free survival (DMFS), tumor-specific survival (TSS), OS and DFS (disease-free survival).

2. Material and methods

The flowchart of image feature extraction, model training and evaluation is illustrated in Fig. 1.

2.1. Patient demographics, imaging data and endpoints

The cohort that was used for model development is the OPC-Radiomics dataset [41] which includes 606 OPSCC patients who received (chemo-) radiotherapy at Princess Margaret Cancer Centre. From this set, 524 patients with planning CT scans (around 75 % were contrast-enhanced) and manually delineated gross tumor volumes of the primary tumor (GTVt) available were randomly split into subsets of n1 = 174, n2 = 200 and n3 = 150 patients that were used for the autoencoder training, building outcome prediction models and independent internal testing, respectively. A detailed description of the OPC-Radiomics set is publicly available at TCIA [41]. The external test (n4 =

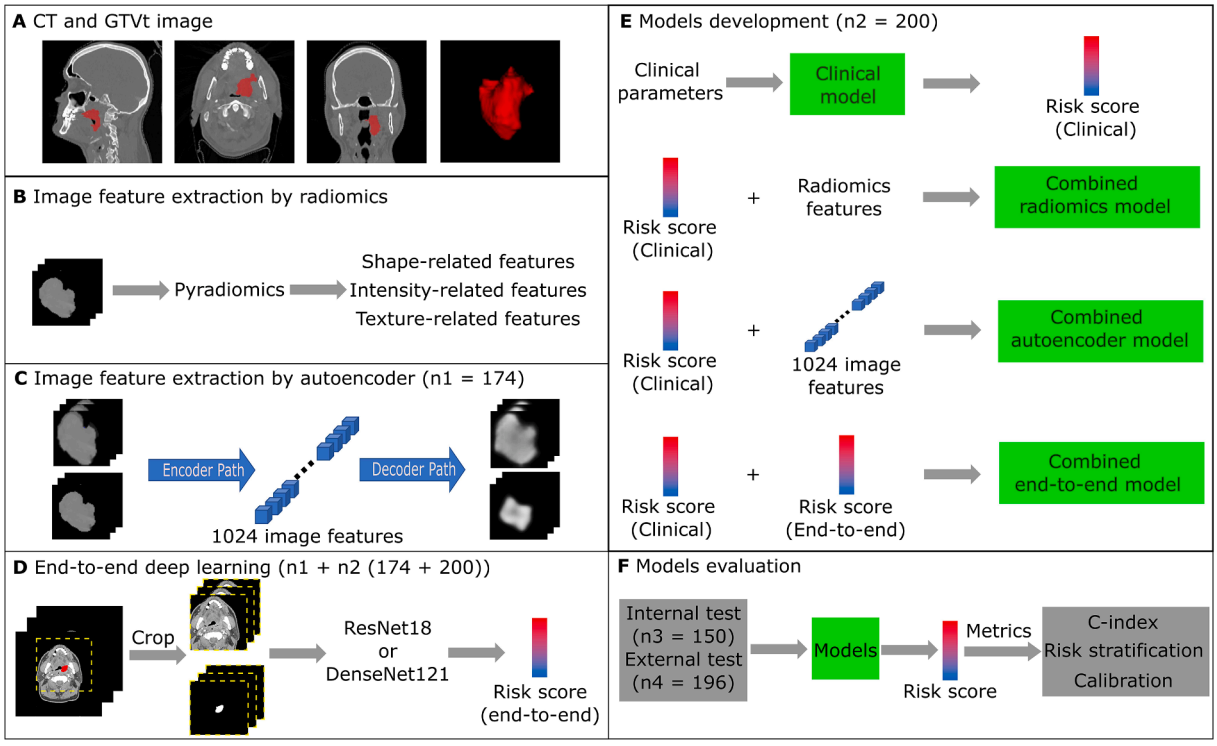

Fig. 1. Flowchart of feature extraction, model training and evaluation of OPSCC outcome prediction models. **A** Example of planning CT scan in three directions and the corresponding gross primary tumor volume (GTVt). **B** Image feature extraction by radiomics. **C** Autoencoder for extracting image features from CT tumor images. **D** End-to-end models inputting CT and GTVt for risk score prediction. **E** Models development using Cox regression. **F** Models evaluation.

197) was performed on the UMCG-OPC dataset (detailed description in Supplementary 1) comprising 197 OPSCC patients. All OPC-Radiomics and UMCG-OPC patients were without distant metastasis at diagnosis.

Clinical candidate predictors and outcome endpoints are explained in Supplementary 2.

2.2. Image feature extraction by radiomics

In Fig. 1B, 110 radiomics features were extracted for each patient (details in Supplementary 3).

2.3. Image feature extraction by autoencoder

Image features were extracted from the GTVt volume in the CT by the pyramid autoencoder, which is a self-supervised learning-based CNN. The input has two channels which were obtained by the method described in Supplementary 4. The autoencoder (Fig. 1C) consists of an encoder path for extracting 1024 representative image features from the input 3D CT tumor images and a decoder aiming to reconstruct the input images from the image features. The detailed description of the architecture (Figure S1) and training strategies of the autoencoder are displayed in Supplementary 4. Each autoencoder extracted image feature was normalized by the maximum and minimum values of this image feature in the training cohort.

2.4. End-to-end deep learning

An end-to-end deep learning method was built to directly predict the risk score of each endpoint (Fig. 1D, detailed description in Supplementary 5).

2.5. Outcome prediction models development

From Fig. 1E, clinical models were first built using multivariable Cox proportional hazard regression analysis for the prediction of each outcome endpoint. The forward selection process was repeated 1000 times using bootstrapping samples in the patients from n1 and n2. In each time of forward selection, the C-index increase was the criteria and the largest number of selected features was set to 5. After each forward selection, only significant predictors ($p < 0.05$ of the Wald test) were finally selected. After 1000 times of forward selection, the selected frequency of features in all 1000 times were ranked. Then, the most frequently selected clinical features which had a significant contribution ($p < 0.05$ of the Wald test) when used together in the clinical model were included in the final clinical model.

For the combined radiomics and autoencoder models, the linear predictor of the clinical model was considered as one feature together with the radiomics features or the 1024 image features extracted by the autoencoder and were used to perform the same multivariable Cox proportional hazard regression analysis, as shown for the clinical model above. Then, the most frequently selected features were used to build combined radiomics and autoencoder models.

The two predicted risk scores from the clinical model and the end-to-end deep learning model were used to build combined end-to-end models for each endpoint. The models implementation code can be found in https://github.com/baoqiangmaUMCG/Ctfeatures-Outcome-Prediction.

2.6. Model performance evaluation

The concordance index (C-index) [95 % confidence interval (CI)] was first applied to evaluate the discriminative ability, with the z-test to compare C-index differences between models. In detail, the z-test compared the difference of 1000 C-indexes calculated on 1000 bootstrapping samples. Then, patients were divided into a high-risk group (hazard value > the median hazard of the training set) and a low-risk group (hazard value <= the median) for each endpoint, with the log-rank tests [42] to determine the significant differences of Kaplan-Meier (KM) curves between groups for each outcome. Additionally, the calibration ability of the models was determined by comparing the curves of outcome rates predicted by the combined models with the KM curves (95 % CI) of the observed outcomes within 5-year follow-up. The predicted curves of each outcome endpoint such as OS were obtained by averaging the predicted OS curves of each patient in the test cohorts. Finally, we compared actual and predicted 2-year outcomes rate with the Hosmer Lemeshow (HS) test for evaluating goodness of fit and calculated the calibration slope and intercept. A two-tailed p-value < 0.05 was considered significant. A significant HS test indicates a poor calibration.

3. Results

Table S1 and Table S2 displayed the differences between OPC-Radiomics and UMCG OPC in clinical data and outcomes (details in Supplementary 6).

Fig. 2 displays four examples of the input and output (reconstructed by the autoencoder) CT tumor images. The shapes and CT intensities of tumors were generally reconstructed by the autoencoder, which means that the extracted image features are representative and descriptive for the tumor images. Additionally, the autoencoder achieved the mean squared error of 0.037 and 0.020, and the Structural Similarity [43] of 0.704 and 0.706, in the internal and external tests, respectively.

The C-index values of clinical, combined autoencoder and combined radiomics models are shown in Table 1, in which the clinical models achieved high C-index values in the training set (between 0.67 and 0.81), the independent internal test set (from 0.60 to 0.76) and the external test set (from 0.67 to 0.80). The combined autoencoder models obtained higher C-index values than clinical models for all outcomes in the training and independent internal test sets, with the largest C-index improvements in RC (from 0.76 to 0.91), DMFS (from 0.60 to 0.74) and the smallest improvement < 0.01 for OS in the independent internal test set. The C-index of 0.74 for DMFS is comparable to 0.69, the best test C-index in OPC-radiomics set, achieved by the CNN models proposed by Lombardo et. al [39]. In the external test set, combined autoencoder models achieved higher C-index values in LC, LRC, TSS (C-index improvement < 0.01), OS (C-index improvement < 0.01) and DFS than clinical models with the highest C-index improvement from 0.71 to 0.76 in LC. The combined radiomics models achieved higher C-index values than clinical models for all endpoints in the training and external test sets, and for LC, RC, LRC, DMFS, OS (improvement < 0.01) in the internal test set. After comparison of the three models, the combined autoencoder models obtained significant highest C-indexes for all endpoints except OS in the internal test set while the combined radiomics models had the highest C-indexes for all endpoints in the external test set. As displayed in Table S4, the combined end-to-end models did not achieve significantly higher C-index values for most endpoints in both internal and external test sets while it did in the training set. The combined models generally keep their better performance than clinical models for HPV positive and negative patients, respectively (Supplementary 8).

Fig. 3 and Figure S2 show the KM curves of high and low risk groups stratified by the clinical (A), combined autoencoder models (B) and combine radiomics models (C) for LC, DMFS and OS, and all endpoints, respectively. The p-values of the log rank tests show that clinical models can stratify patients with significant differences for (LRC, TSS, OS and DFS) and (RC, LRC, TSS, OS and DFS) in the internal and external test sets, respectively. Combined autoencoder models and combined radiomics models showed significant differences for all endpoints except for DMFS in the external test set and except for DMFS in the internal test set, respectively.

Fig. 4A and Figure S3A show the calibration curves of the combined autoencoder models. The 95 % CIs of predicted curves and the actual KM

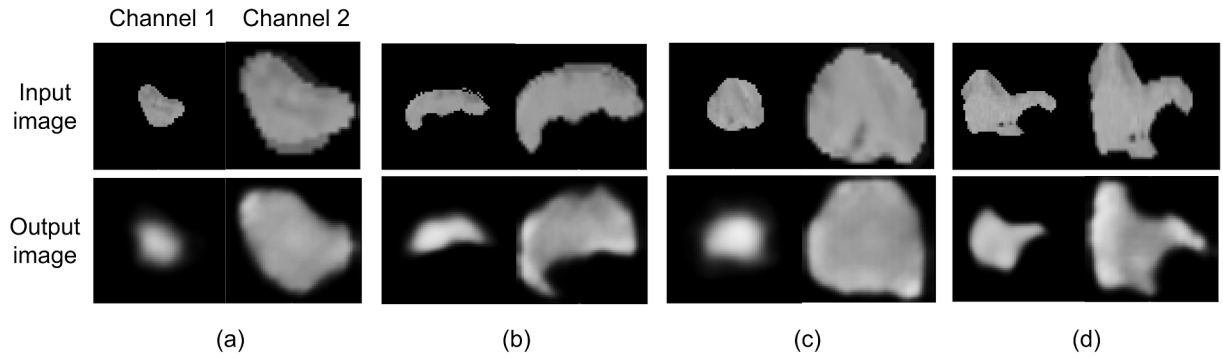

Fig. 2. Four examples of input and output CT tumor images of autoencoders.

Table 1
C-index [95% confidence interval] results of clinical and combined autoencoder or radiomics models.

	Training set			Internal test set			External test set		
	Clinical model	Combined autoencoder model	Combined radiomics model	Clinical model	Combined autoencoder model	Combined radiomics model	Clinical model	Combined autoencoder model	Combined radiomics model
LC	0.81 [0.73,0.89]	**0.84** **[0.75,0.92]**	0.84 [0.74,0.92]	0.73 [0.55,0.85]	**0.79** **[0.59,0.93]** *	0.79[0.57, 0.92]	0.71 [0.56,0.86]	0.76[0.64,0.86]	**0.82** **[0.72,0.90]** *
RC	0.74 [0.62,0.86]	**0.86** **[0.76,0.93]** *	0.79 [0.67,0.91]	0.76 [0.52,0.96]	**0.91** **[0.82,0.98]** *	0.84 [0.70,0.97]	0.70 [0.57,0.81]	0.65[0.52,0.76]	**0.76** **[0.67,0.85]** *
LRC	0.71 [0.60,0.81]	**0.81** **[0.74,0.88]** *	0.74 [0.63,0.85]	0.72 [0.59,0.85]	**0.78** **[0.63,0.91]** *	0.77 [0.66,0.88]	0.71 [0.62,0.79]	0.72[0.64,0.81]	**0.75** **[0.66,0.83]** *
DMFS	0.67 [0.58,0.75]	**0.75** **[0.66,0.83]** *	0.73 [0.65,0.81]	0.60 [0.49,0.72]	**0.74** **[0.63,0.84]** *	0.67 [0.55,0.78]	0.67 [0.58,0.76]	0.64[0.50,0.78]	**0.72** **[0.61,082]** *
TSS	0.76 [0.67,0.83]	**0.79** **[0.72,0.85]** *	0.77 [0.70,0.84]	0.70 [0.59,0.80]	**0.71** **[0.60,0.81]** *	0.70 [0.59,0.80]	0.77 [0.68,0.85]	0.77[0.68,0.85]	**0.79** **[0.69, 0.87]** *
OS	0.73 [0.68,0.78]	**0.76** **[0.71,0.80]** *	0.75[0.70, 0.79]	0.72 [0.64,0.78]	0.72[0.64,0.78]	**0.72** **[0.64, 0.78]**	0.80 [0.74,0.86]	0.80[0.74,0.86]	**0.81** **[0.75,0.87]** *
DFS	0.71 [0.66,0.76]	**0.74** **[0.69,0.79]**	0.74 [0.68,0.79]	0.72 [0.65,0.78]	**0.73** **[0.66,0.80]** *	0.72 [0.64,0.78]	0.71 [0.64,0.79]	0.73[0.66,0.80]	**0.77** **[0.71,0.83]** *

The underlined C-indexes are slightly higher (C-index improvement < 0.01) than that of the other model in the same dataset. *: Significant difference of C-indexes between the clinical model and combined model (p-Value < 0.05 by z-test).

curves overlap for LC, DMFS, TSS, OS and DFS in the internal test set and for LC, RC, LRC, DMFS and DFS in the external test set within a 2 year follow up period. According to the p-values in Fig. 4B and S3B, combined autoencoder models showed good calibration performance ($p > 0.05$ by HS test) for 2-year LC, DMFS, TSS, OS and DFS in the internal test set and for all 2-year endpoints except RC in the external test set. Additionally, the obtained real 2-year calibration lines are good (slope within [0.8, 1.2] and intercept within [-0.2, 0.2]) for all endpoints except RC in the external test. Curves of combined radiomics models are descripted in Supplementary 9.

4. Discussion

This study investigated and compared the added value of CT-image features extracted by radiomics, self-supervised learning (autoencoder) and end-to-end deep learning, respectively in the prediction of LC, RC, LRC, DMFS, TSS, OS and DFS of OPSCC patients treated with (chemo)radiotherapy. Combined autoencoder models which combined self-supervised learning extracted image features with a linear predictor from a clinical prediction model showed better discriminative performance for most outcomes in the internal test set than combined radiomics models. However, combined radiomics models showed best predictive performance for all endpoints in the external test set, which shows the better generalizability of radiomics features than autoencoder extracted features. Combined end-to-end models did not perform better than combined autoencoder models for most endpoints in both the internal and external test sets.

Compared with clinical models, combined autoencoder models achieved higher C-index values for all endpoints in the training and internal test sets as shown in Table 1, which indicates that the self-supervised learning (SLL) extracted high-level CT image features provide complementary information for outcome prediction. Other studies identified SLL extracted image features that were predictive for ovarian cancer [44], colorectal cancer [45] and gastric cancer [46]. In the external test set our combined autoencoder models still performed generally better than clinical models. However, the C-index differences between combined autoencoder and clinical models were small (<0.01) for OS in the internal test set and TSS and OS in the external test set (Table 1). This may be due to that the clinical models using four and five predictors (Table S3) for TSS, and OS prediction, respectively, already achieved a satisfactory performance and image features could not add much new predictive information.

Although combined radiomics models achieved higher C-indexes than clinical models for most endpoints in the internal test set, they were still worse than combined autoencoder models (Table 1). This is most probably because the autoencoder can extract more comprehensive and representative features which can provide more information for outcome prediction than radiomics features. However, in the external test set, combined radiomics models showed significantly higher C-indexes than the combined autoencoder models in all endpoints, which demonstrates that the selected radiomics features (Table S3) have better external generalizability than autoencoder extracted features. From Table S3, we can observe that the combined radiomics models for LC, RC, DMFS, OS and DFS prediction contained mainly shape features that are possibly less affected by differences between CT equipment and scan protocols between institutions than autoencoder extracted features. The

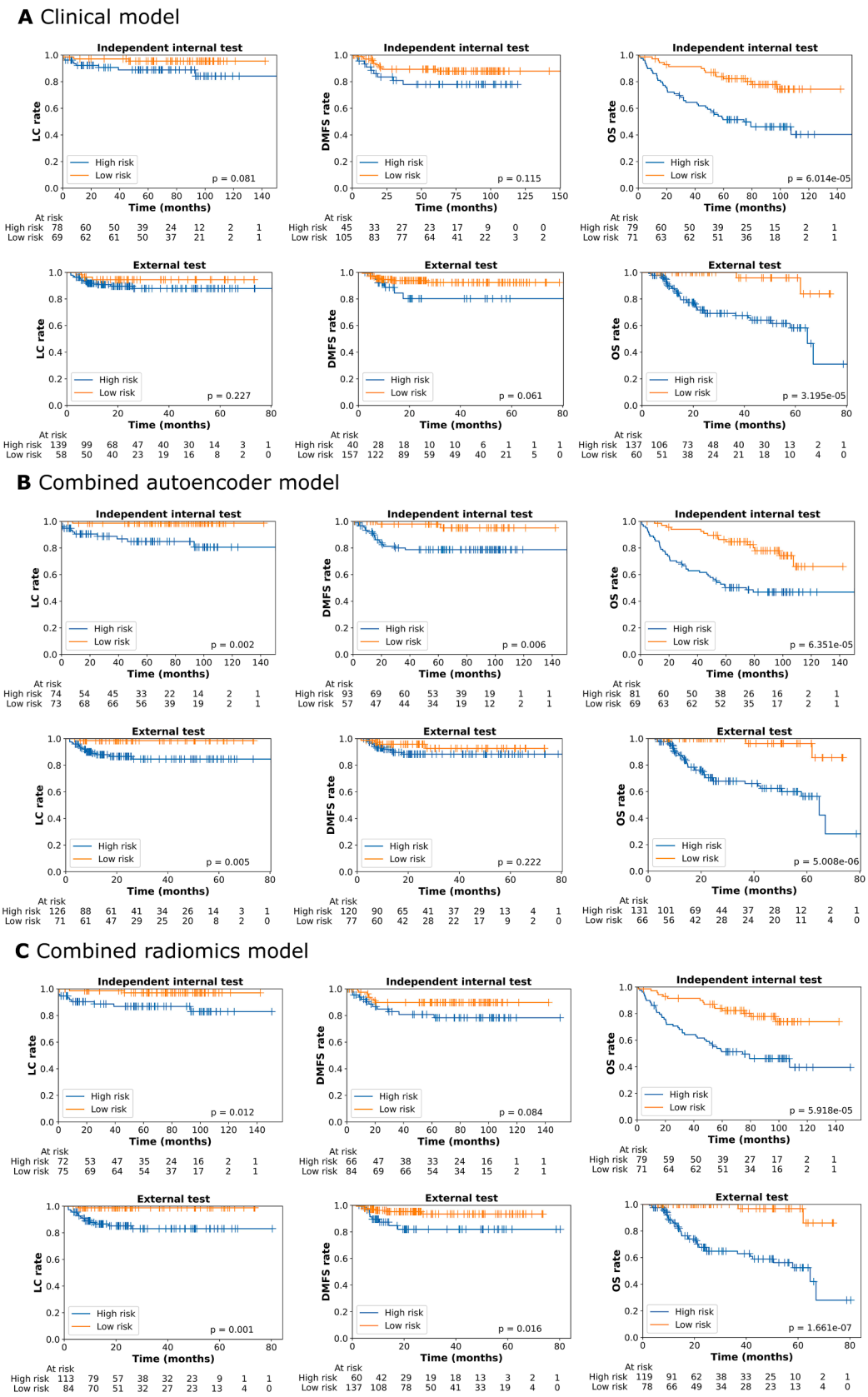

Fig. 3. KM curves of high (hazard values > median) and low (hazard values <= median) risk groups of LC, DMFS and OS in the independent internal and external test sets stratified by clinical models **(A)** and combined autoencoder models **(B)** and combined radiomics models **(C)**. P-values < 0.05 means significant difference between two risk groups.

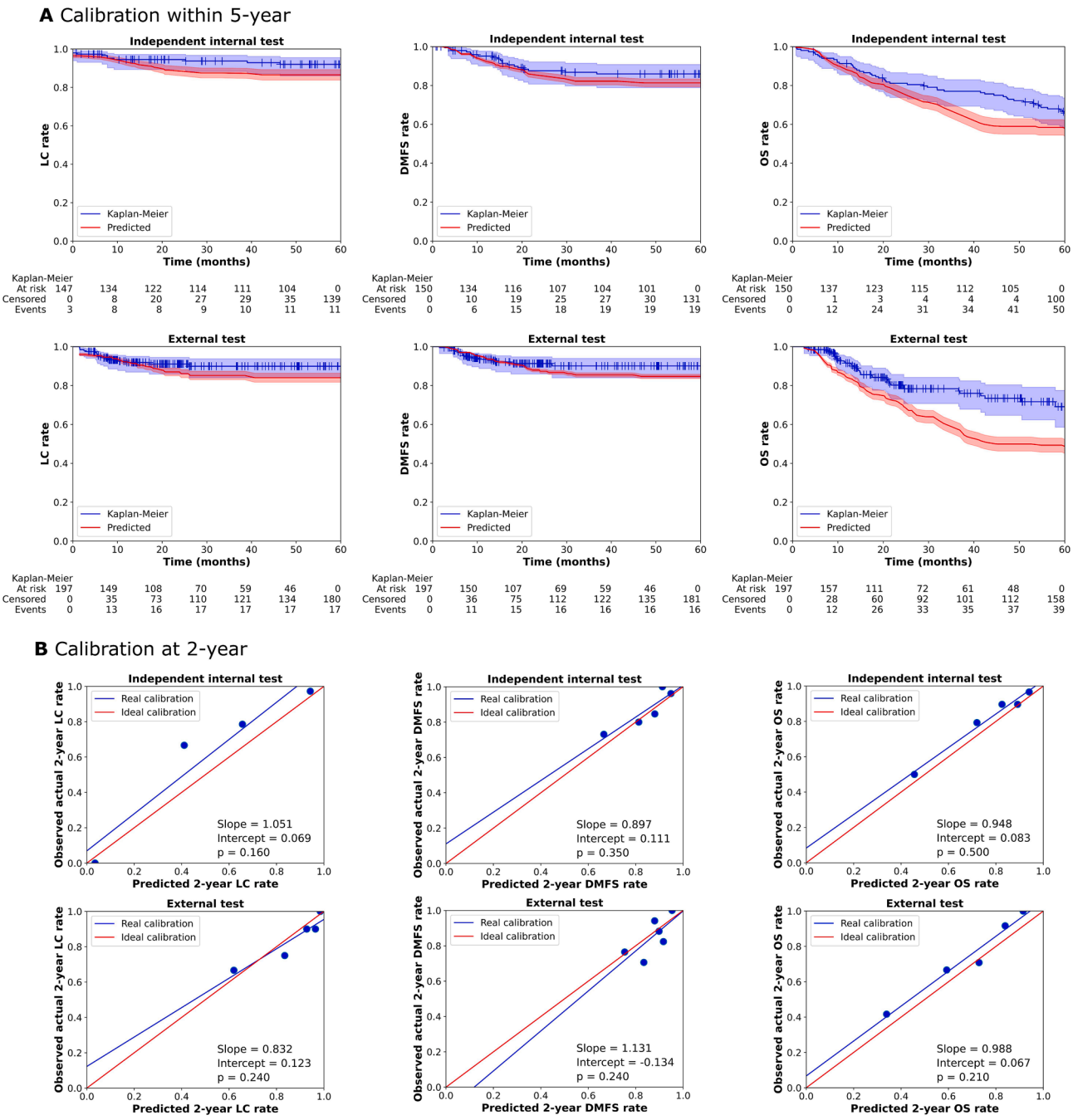

Fig. 4. Calibration performance of combined autoencoder models for LC, DMFS and DFS in the independent internal and external test sets in the (**A**) within 5-year (**B**) at 2-year. p-Values were from HS tests. Slope and intercept belong to real calibration line. The figures in (A) showed the comparison of real Kaplan-Meier curves and the predicted outcome rates curves by combined models within 5-year follow-up.

combined end-to-end models only performed better than the autoencoder combined models for most endpoints in the training set and not in the test sets. This indicates that there was an overfitting problem.

In Fig. 3 and Figure S2, combined (autoencoder or radiomics) models stratified patients into high and low risk groups with significant differences ($p < 0.05$) for LC, RC and DMFS in the independent internal test set and LC in the external test sets while clinical models did not. This demonstrates that combined models could be more powerful in identifying high- and low- risk patients for individualized treatment by adding image features extracted by autoencoder or radiomics. Additionally, we found that the combined radiomics model could achieve a significant risk stratification for DMFS in the external test set while the combined autoencoder model could not. This may demonstrate that the shape feature: original_shape_MinorAxisLength selected for DMFS prediction (Table S3) is more stable than autoencoder extracted features when using it externally.

Combined autoencoder models showed good calibration for most endpoints in both test sets. For example, the combined autoencoder model achieved a good calibration slope of 0.988 and intercept of 0.067 for 2-year OS rate prediction (Fig. 4B) in the external test set as well as higher C-index of 0.80 (Table 1) and better OS risk stratification (Fig. 3) than the clinical model. Similarly, combined autoencoder model showed good calibration (Fig. 4A and 4B), better C-index values (Table 1) and better risk stratification (Fig. 3) than the clinical model for LC rate in both internal and external test sets. Thus, combined autoencoder models are highly effective for LC and OS prediction even in the external test set.

This makes the combined autoencoder models promising for clinical tools for selecting patients for personalized treatments. However, in Fig. 4 and Figure S2, the 2-year calibration curves of combined autoencoder models for RC and LRC in the internal test set and RC in the external test set showed a p-value < 0.05 in the HS test indicating a poor calibration. This may be due to the low numbers of 2-year events (RR: 5 and 18 in the internal and external test sets, respectively and LRC: 10 in the internal test set). Combined radiomics models generally showed worse internal test calibration (Supplementary 9).

Additionally, we found that the tumor-volume, a radiomics feature, is not highly related to linear predictors of our combined models, and our results support previous studies [47] that deep learning features are more predictive internally and radiomics features are more stable externally (Supplementary 10). The limitations are shown in Supplementary 11.

In conclusion, we compared the abilities of CT image features extracted by radiomics, self-supervised learning and end-to-end deep learning, respectively, in improving the performance of clinical data-based prediction models for most outcomes in oropharyngeal squamous cell carcinoma patients. Self-supervised extracted features showed better predictive performance in the internal test set while radiomics features showed better generalizability when being used in the external dataset.

CRediT authorship contribution statement

Baoqiang Ma: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Jiapan Guo:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Hung Chu:** Methodology. **Lisanne V. van Dijk:** Supervision, Writing – review & editing. **Peter M.A. van Ooijen:** Supervision, Writing – review & editing. **Johannes A. Langendijk:** Data curation, Funding acquisition, Writing – review & editing. **Stefan Both:** Supervision, Writing – review & editing. **Nanna M. Sijtsema:** Project administration, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Baoqiang gratefully acknowledges the financial support for his PhD study provided by the China Scholarship Council (CSC) and University of Groningen (RUG), and the High Performance Computing cluster provided by RUG.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2023.100502.

References

[1] O'Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. Lancet Oncol 2016:17. https://doi.org/10.1016/S1470-2045(15)00560-4.

[2] Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 2010;363:24–35. https://doi.org/10.1056/NEJMoa0912217.

[3] Bos P, van den Brekel MWM, Gouw ZAR, Al-Mamgani A, Taghavi M, Waktola S, et al. Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. Eur J Radiol 2021:139. https://doi.org/10.1016/j.ejrad.2021.109701.

[4] Kühn JP, Schmid W, Körner S, Bochen F, Wemmert S, Rimbach H, et al. Hpv status as prognostic biomarker in head and neck cancer—which method fits the best for outcome prediction? Cancers (Basel) 2021:13. https://doi.org/10.3390/cancers13184730.

[5] Yin LX, D'Souza G, Westra WH, Wang SJ, van Zante A, Zhang Y, et al. Prognostic factors for human papillomavirus–positive and negative oropharyngeal carcinomas. Laryngoscope 2018:128. https://doi.org/10.1002/lary.27130.

[6] de França GM, da Silva WR, Medeiros CKS, Júnior JF, de Moura SE, Galvão HC. Five-year survival and prognostic factors for oropharyngeal squamous cell carcinoma: retrospective cohort of a cancer center. Oral Maxillofac Surg 2021. https://doi.org/10.1007/s10006-021-00986-4.

[7] Fakhry C, Zhang Q, Nguyen-Tân PF, Rosenthal DI, Weber RS, Lambert L, et al. Development and validation of nomograms predictive of overall and progression-free survival in patients with oropharyngeal cancer. J Clin Oncol 2017;35. https://doi.org/10.1200/JCO.2016.72.0748.

[8] Rios Velazquez E, Hoebers F, Aerts HJWL, Rietbergen MM, Brakenhoff RH, Leemans RC, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Radiother Oncol 2014:113. https://doi.org/10.1016/j.radonc.2014.09.005.

[9] Chen SY, Last A, Ettyreddy A, Kallogjeri D, Wahle B, Chidambaram S, et al. 20 pack-year smoking history as strongest smoking metric predictive of HPV-positive oropharyngeal cancer outcomes. Am J Otolaryngol - Head Neck Med Surg 2021:42. https://doi.org/10.1016/j.amjoto.2021.102915.

[10] Ward MJ, Thirdborough SM, Mellows T, Riley C, Harris S, Suchak K, et al. Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer. Br J Cancer 2014:110. https://doi.org/10.1038/bjc.2013.639.

[11] Xiao R, Pham Y, Ward MC, Houston N, Reddy CA, Joshi NP, et al. Impact of active smoking on outcomes in HPV+ oropharyngeal cancer. Head Neck 2020:42. https://doi.org/10.1002/hed.26001.

[12] Langius JAE, Bakker S, Rietveld DHF, Kruizenga HM, Langendijk JA, Weijs PJM, et al. Critical weight loss is a major prognostic indicator for disease-specific survival in patients with head and neck cancer receiving radiotherapy. Br J Cancer 2013:109. https://doi.org/10.1038/bjc.2013.458.

[13] Zhang Q, Zhou Z, Qin G, Li P, Sher DJ, Wang J, et al. Prediction of Local Persistence/Recurrence on PET/CT scans after Radiation Therapy Treatment of Head and Neck Cancer Using a Multi-objective Radiomics Model. Int. J Radiat Oncol 2018:102. https://doi.org/10.1016/j.ijrobp.2018.06.243.

[14] Haider SP, Zeevi T, Baumeister P, Reichel C, Sharaf K, Forghani R, et al. Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma. Cancers (Basel) 2020;12. https://doi.org/10.3390/cancers12071778.

[15] Haider SP, Sharaf K, Zeevi T, Baumeister P, Reichel C, Forghani R, et al. Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics. Transl Oncol 2021:14. https://doi.org/10.1016/j.tranon.2020.100906.

[16] Leijenaar RTH, Carvalho S, Hoebers FJP, Aerts HJWL, Van Elmpt WJC, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol (Madr) 2015:54. https://doi.org/10.3109/0284186X.2015.1061214.

[17] Rich B, Huang J, Yang Y, Jin W, Johnson P, Wang L, et al. Radiomics predicts for distant metastasis in locally advanced human papillomavirus-positive oropharyngeal squamous cell carcinoma. Cancers (Basel) 2021:13. https://doi.org/10.3390/cancers13225689.

[18] Ma B, Zhao Y, Yang Y, Zhang X, Dong X, Zeng D, et al. MRI image synthesis with dual discriminator adversarial learning and difficulty-aware attention mechanism for hippocampal subfields segmentation. Comput Med Imaging Graph 2020:86. https://doi.org/10.1016/j.compmedimag.2020.101800.

[19] Zhao Y, Ma B, Jiang P, Zeng D, Wang X, Li S. Prediction of Alzheimer's Disease Progression with Multi-Information Generative Adversarial Network. IEEE J Biomed Heal Informatics 2021:25. https://doi.org/10.1109/JBHI.2020.3006925.

[20] Zhao Y, Ma B, Che T, Li Q, Zeng D, Wang X, et al. Multi-view prediction of Alzheimer's disease progression with end-to-end integrated framework. J Biomed Inform 2022:125. https://doi.org/10.1016/j.jbi.2021.103978.

[21] Zhang X, Kelkar VA, Granstedt J, Li H, Anastasio MA. Impact of deep learning-based image super-resolution on binary signal detection. J Med Imaging 2021:8. https://doi.org/10.1117/1.jmi.8.6.065501.

[22] Kelkar VA, Zhang X, Granstedt J, Li H, Anastasio MA. Task-based evaluation of deep image super-resolution in medical imaging, 2021. https://doi.org/10.1117/12.2582011.

[23] Zeng D, Li Q, Ma B, Li S. Hippocampus segmentation for preterm and aging brains using 3D densely connected fully convolutional networks. IEEE Access 2020:8. https://doi.org/10.1109/ACCESS.2020.2993504.

[24] De Biase A, Tang W, Sourlos N, Ma B, Guo J, Sijtsema NM, et al. Skip-SCSE Multi-scale Attention and Co-learning Method for Oropharyngeal Tumor Segmentation on Multi-modal PET-CT Images, 2022. https://doi.org/10.1007/978-3-030-98253-9_10.

[25] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiat Oncol Biol Phys 2019:104. https://doi.org/10.1016/j.ijrobp.2019.02.040.

[26] Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. Int J Radiat Oncol Biol Phys 2018:101. https://doi.org/10.1016/j.ijrobp.2018.01.114.

[27] Mao X, Pineau J, Keyes R, Enger SA. RapidBrachyDL: rapid radiation dose calculations in brachytherapy via deep learning. Int J Radiat Oncol Biol Phys 2020:108. https://doi.org/10.1016/j.ijrobp.2020.04.045.
[28] Cui S, Ten Haken RK, El Naqa I. Integrating multiomics information in deep learning architectures for joint actuarial outcome prediction in non-small cell lung cancer patients after radiation therapy. Int J Radiat Oncol Biol Phys 2021:110. https://doi.org/10.1016/j.ijrobp.2021.01.042.
[29] Jiao Z, Li H, Xiao Y, Dorsey J, Simone CB, Feigenberg S, et al. Integration of deep learning radiomics and counts of circulating tumor cells improves prediction of outcomes of early stage NSCLC patients treated with stereotactic body radiation therapy. Int J Radiat Oncol Biol Phys 2022:112. https://doi.org/10.1016/j.ijrobp.2021.11.006.
[30] Ma B, Guo J, Zhai T, van der Schaaf A, Steenbakkers RJHM, van Dijk LV, et al. CT-based deep multi-label learning prediction model for outcome in patients with oropharyngeal squamous cell carcinoma. Med Phys 2023.
[31] Fujima N, Andreu-Arasa VC, Meibom SK, Mercier GA, Truong MT, Hirata K, et al. Prediction of the local treatment outcome in patients with oropharyngeal squamous cell carcinoma using deep learning analysis of pretreatment FDG-PET images. BMC Cancer 2021:21. https://doi.org/10.1186/s12885-021-08599-6.
[32] Cheng NM, Yao J, Cai J, Ye X, Zhao S, Zhao K, et al. Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using FDG-PET imaging. Clin Cancer Res 2021:27. https://doi.org/10.1158/1078-0432.CCR-20-4935.
[33] Wahid KA, He R, Dede C, Mohamed ASR, Abdelaal MA, van Dijk LV, et al. Combining tumor segmentation masks with PET/CT images and clinical data in a deep learning framework for improved prognostic prediction in head and neck squamous cell carcinoma. MedRxiv 2021.
[34] Ma B, Guo J, De BA, Sourlos N, Tang W, van Ooijen P, et al. Self-supervised multi-modality image feature extraction for the progression free survival prediction in head and neck cancer. In: 3D Head Neck Tumor Segmentation PET/CT Chall. Springer; 2021. p. 308–17.
[35] Ma B, Li Y, Chu H, Tang W, De la O Arévalo LR, Guo J, et al. Deep learning and radiomics based PET/CT image feature extraction from auto segmented tumor volumes for recurrence-free survival prediction in oropharyngeal cancer patients, 2023. https://doi.org/10.1007/978-3-031-27420-6_24.
[36] Ma B, Guo J, Van Dijk L, van Ooijen PMA, Both S, Sijtsema NM. TransRP: Transformer-based PET/CT feature extraction incorporating clinical data for recurrence-free survival prediction in oropharyngeal cancer. Med Imaging with Deep Learn 2023.
[37] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction n.d. https://doi.org/10.1038/s41598-019-39206-1.
[38] Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 2017:7. https://doi.org/10.1038/s41598-017-10371-5.
[39] Lombardo E, Kurz C, Marschner S, Avanzo M, Gagliardi V, Fanetti G, et al. Distant metastasis time to event analysis with CNNs in independent head and neck cancer cohorts. Sci Rep 2021;11:6418.
[40] Wang Y, Lombardo E, Avanzo M, Zschaek S, Weingärtner J, Holzgreve A, et al. Deep learning based time-to-event analysis with PET, CT and joint PET/CT for head and neck cancer prognosis. 106948 Comput Methods Programs Biomed 2022.
[41] Kwan JYY, Su J, Huang S, Ghoraie L, Xu W, Chan B, et al. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma. Cancer Imaging Arch 2019.
[42] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50.
[43] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004:13. https://doi.org/10.1109/TIP.2003.819861.
[44] Wang S, Liu Z, Rong Y, Zhou B, Bai Y, Wei W, et al. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. Radiother Oncol 2016. https://doi.org/10.1016/j.radonc.2018.10.019.
[45] Zhao J, Wang H, Zhang Y, Wang R, Liu Q, Li J, et al. Deep learning radiomics model related with genomics phenotypes for lymph node metastasis prediction in colorectal cancer. Radiother Oncol 2022;167:195–202.
[46] Zhang W, Fang M, Dong D, Wang X, Ke X, Zhang L, et al. Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. Radiother Oncol 2020:145. https://doi.org/10.1016/j.radonc.2019.11.023.
[47] Chen J, Wee L, Dekker A, Bermejo I. Using 3D deep features from CT scans for cancer prognosis based on a video classification model: A multi-dataset feasibility study. Med Phys 2023.