Deep learning-based outcome prediction using PET/CT and automatically predicted probability maps of primary tumor in patients with oropharyngeal cancer

De Biase, Alessia; Ma, Baoqiang; Guo, Jiapan; van Dijk, Lisanne V; Langendijk, Johannes A; Both, Stefan; van Ooijen, Peter M A; Sijtsema, Nanna M

Published in:
Computer Methods and Programs in Biomedicine

DOI:
10.1016/j.cmpb.2023.107939

Computer Methods and Programs in Biomedicine 244 (2024) 107939

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Deep learning-based outcome prediction using PET/CT and automatically predicted probability maps of primary tumor in patients with oropharyngeal cancer

Alessia De Biase [a,b,1], Baoqiang Ma [a,*,1], Jiapan Guo [c], Lisanne V. van Dijk [a], Johannes A. Langendijk [a], Stefan Both [a], Peter M.A. van Ooijen [a,b], Nanna M. Sijtsema [a]

[a] Department of Radiation Oncology, University Medical Centre Groningen (UMCG), RB, Groningen 9700, the Netherlands
[b] Data Science Centre in Health (DASH), University Medical Centre Groningen (UMCG), RB, Groningen 9700, the Netherlands
[c] Computer Science and Artificial Intelligence, Bernoulli Institute for Mathematics, University of Groningen (RUG), Groningen, AK 9700, the Netherlands

ARTICLE INFO

Keywords:
Deep learning
PET/CT
Oropharyngeal cancer
Outcome prediction
Tumor probability map

ABSTRACT

Background and Objective: Recently, deep learning (DL) algorithms showed to be promising in predicting outcomes such as distant metastasis-free survival (DMFS) and overall survival (OS) using pre-treatment imaging in head and neck cancer. Gross Tumor Volume of the primary tumor (GTVp) segmentation is used as an additional channel in the input to DL algorithms to improve model performance. However, the binary segmentation mask of the GTVp directs the focus of the network to the defined tumor region only and uniformly. DL models trained for tumor segmentation have also been used to generate predicted tumor probability maps (TPM) where each pixel value corresponds to the degree of certainty of that pixel to be classified as tumor. The aim of this study was to explore the effect of using TPM as an extra input channel of CT- and PET-based DL prediction models for oropharyngeal cancer (OPC) patients in terms of local control (LC), regional control (RC), DMFS and OS.

Methods: We included 399 OPC patients from our institute that were treated with definitive (chemo)radiation. For each patient, CT and PET scans and GTVp contours, used for radiotherapy treatment planning, were collected. We first trained a previously developed 2.5D DL framework for tumor probability prediction by 5-fold cross validation using 131 patients. Then, a 3D ResNet18 was trained for outcome prediction using the 3D TPM as one of the possible inputs. The endpoints were LC, RC, DMFS, and OS. We performed 3-fold cross validation on 168 patients for each endpoint using different combinations of image modalities as input. The final prediction in the test set (100) was obtained by averaging the predictions of the 3-fold models. The C-index was used to evaluate the discriminative performance of the models.

Results: The models trained replacing the GTVp contours with the TPM achieved the highest C-indexes for LC (0.74) and RC (0.60) prediction. For OS, using the TPM or the GTVp as additional image modality resulted in comparable C-indexes (0.72 and 0.74).

Conclusions: Adding predicted TPMs instead of GTVp contours as an additional input channel for DL-based outcome prediction models improved model performance for LC and RC.

1. Introduction

The most common pre-treatment imaging modalities for patients diagnosed with oropharyngeal cancer (OPC) are computed tomography (CT), magnetic resonance imaging (MRI), and FDG positron emission tomography (PET). Quantitative imaging features derived from pre-treatment imaging modalities demonstrated to be strong prognostic factors in head and neck (H&N) cancers patients [1]. Recent studies have shown that single-modality DL approaches are able to outperform traditional radiomic frameworks in predicting distant metastasis (DM) [2] and overall survival (OS) [3]. Moreover, because of the complementary roles of different imaging modalities in the evaluation and treatment planning of OPC patients, a multi-modality approach could increase the predictive power of outcome prediction models.

* Corresponding author.
E-mail address: b.ma@umcg.nl (B. Ma).
[1] These authors contributed equally.

https://doi.org/10.1016/j.cmpb.2023.107939
Received 28 February 2023; Received in revised form 20 November 2023; Accepted 20 November 2023
Available online 22 November 2023

In previous studies, cropping or masking the patient scans with the tumor manual delineation were two alternative techniques involved in the pre-processing steps of DL outcome prediction models [4–6]. This allows the network to focus on the regions where relevant information is present, reducing the model training complexity. However, peri-tumoral regions, such as adjacent tissue invasion, have potential for predicting outcome [7–9]. Hence, it may be beneficial to include a larger volume than the tumor region only [10]. Thus, there is an unmet need for removing the non-relevant background information, while not limiting the tumor region of the DL input excessively. Including the Gross Tumor Volume (GTV) segmentation as an additional channel to the input data was suggested, in some studies [2,11,12], to improve the performance of outcome prediction models. The use of a binary mask as additional input channel guides deep learning networks in learning tumor shape while enforcing a clear differentiation between GTV and the surrounding background.

Multi-task frameworks performing both segmentation and outcome prediction have also been widely explored in recent years, showing that features learned on another task could represent additional prognostic information for survival analysis [13–15]. Andrearczyk et al. were the first ones to show that multi-task networks can apply the knowledge acquired during tumor volume segmentation in PET/CT images to make predictions regarding patient prognosis in H&N cancer [13]. In their work, a 3D UNet was used for GTVp segmentation. Additionally, information collected by skip connections from the downsampling path was employed by fully-connected layers to predict disease-free survival (DFS). Meng et al. [14] introduced a 3D end-to-end Deep Multi-Task Survival model (DeepMTS) that simultaneously performs tumor segmentation and survival prediction using PET/CT of nasopharyngeal cancer patients. Compared to the previously cited work, the branch of

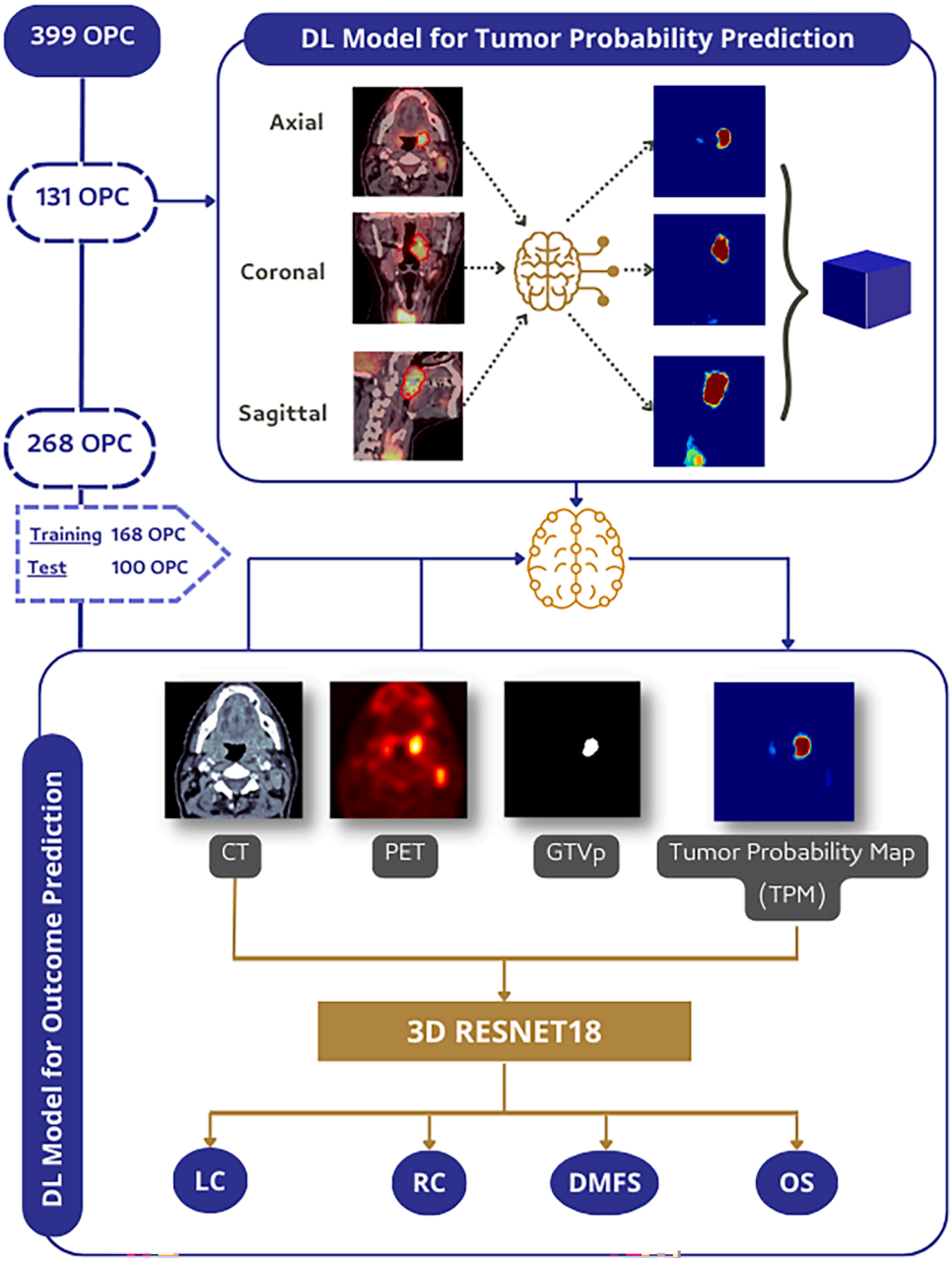

Fig. 1. Study design and deep learning framework. Firstly, the DL model for tumor segmentation was trained and validated on PET-CT images of 131 OPC patients. Then, the model was tested on PET-CT images of 268 OPC patients to obtain tumor probability maps. Finally, the DL model for outcome prediction was trained, validated, and tested using different combinations of imaging modalities as input (see first row in the second block).

the network performing outcome prediction is more complex. A cascaded survival network (CSN) extracts deep features from the output of the segmentation model concatenated to the PET/CT input. Furthermore, the latter and the deep features derived from the segmentation backbone are then fed to fully connected layers for survival prediction. The same method was also employed for progression-free survival (PFS) [15] and recurrence-free survival (RFS) [16] prediction in patients with H&N cancer. Although the performance of the DeepMTS were promising, in their latest work the authors mentioned that the DL model did not fully capture the prognostic information in tumor regions. One hypothesis is that intra- and inter- observer variability could still affect the automatic segmentation accuracy [17], which feeds misleading information to prediction models.

In order to capture the model uncertainty in the segmentation task, in our previous study [18], we designed a novel DL model for automated tumor segmentation on registered FDG PET/CT images, which uses spatial and model uncertainty to generate predicted tumor probability maps (TPMs) instead of binary outputs. Therefore, the predicted output is not a fixed segmentation, but rather a probability per voxel, indicating how likely that voxel represents tumor tissue. Tumor probability maps are not confined to defining a region of interest; they are expected to offer a confidence-weighted indication of the segmentation model's prediction for each area being part of the tumor.

The goal and novelty of our current study was to investigate the potential of using predicted tumor probability maps from De Biase et al. [18] as input for DL CT- and PET-based local control (LC), regional control (RC), distant metastasis-free survival (DMFS), and overall survival (OS) prediction models for oropharyngeal cancer (OPC) patients. For each endpoint, the aim was to train and test models using different combinations of input modalities including PET, CT, GTV of the primary tumor (GTVp), and TPM and to analyze their differences in performance (see Fig. 1).

2. Materials and methods

2.1. Data description

We used data of oropharyngeal squamous cell carcinoma (OPSCC) patients who were treated with radiotherapy with/without chemotherapy at the department of Radiation Oncology of the University Medical Center Groningen (UMCG) between 2010 and 2021. All patients were included in a prospective data registration program (Clinical Trials NCT 02,435,576) as part of routine clinical practice, collecting baseline characteristics, toxicity, patient-reported outcome measures, and tumor follow up. A total of 399 OPSCC patients satisfied the inclusion criteria (see B.1).

PET and CT scans, and GTVp delineations used for radiotherapy treatment planning were collected. Details on image acquisition and tumor contour delineation can be found in previous studies [19,20]. Most patients were scanned on a combined PET-CT scanner. The images of patients that were scanned on separate PET- and CT-scanners were registered rigidly using the software Mirada DBx from Mirada Medical Ltd. All imaging was acquired in treatment position using fixation with a mask. The clinical data that was considered as potential predictors is described in B.2.

Outcome endpoints consist of local control (LC), regional control (RC), distant metastasis-free survival (DMFS), and overall survival (OS). The events of LC, RC were defined as residual or recurrent lesions at the primary location and the regional nodes, respectively. Distant metastasis, death due to any reasons are the events of DMFS and OS, respectively. Details can be found in the Appendix B.

2.2. Data pre-processing

A bounding box around the oropharynx was automatically determined for each patient with the method used by De Biase et al. [20]. Then, CT and PET images within the bounding box region were resampled to $1 \times 1 \times 1$ mm^3. CT intensities were normalized by a min-max normalization following truncating pixel values to $[-200, 200]$ to focus on soft tissue contrast only. PET intensities were first truncated to [0, 25] and subsequently normalized to [0, 1] by the min-max method.

2.3. Tumor probability prediction model

Patients treated after 2019 (131 in total) were used to train and validate the model for tumor probability prediction from [18] (Fig. 1). The DL network trains on sequences defined as three consecutive slices extracted by a same orthogonal cross-section of a volume. The output of the network was predicted sequences containing overlapping slices, which were then averaged. 5-fold cross validation was performed three times using sequences extracted from the axial, coronal and sagittal view of concatenated PET-CT images. For each run, the network was trained for 150 epochs and the model was saved at the lowest validation loss value reached after training for 100 epochs. Patients treated before 2019 (268 in total) were used as an independent test set. For each patient in the test set, we performed both single-view average based ensembling and 3D volume reconstruction from the 2D slices. Finally, multi-view average based ensembling was performed fusing the three volumes obtained from the axial, coronal and sagittal view predictions. Refer to De Biase et al. [18] for details on the DL-based method used for tumor probability prediction and hyperparameters.

The reasons behind the data split choice were the following: (1) long follow up was not available for patients after 2019; (2) after a consultation with our radiation oncologists, we concluded that most recent patients had more reliable tumor contours thanks to the incorporation of MRI imaging in the contouring process and overall improved imaging techniques; (3) we believed that maintaining two distinct groups, one for segmentation and another for outcome prediction, would ensure that the probability maps used in training and testing the outcome prediction models were derived from separate and independent datasets.

2.4. Outcome prediction model

2.4.1. Model architecture

Our outcome prediction model in Fig. 2 was constructed based on ResNet [21] which consists of one 3D convolution layer, 8 residual blocks, a global average pooling (GAP) layer and one dense layer. Batch normalization and ReLU activation were used following the first convolution layer and in the residual blocks as shown in Fig. 2. Residual block 1 and residual block 2 perform residual connection by an average pooling with a stride of 2 and a $1 \times 1 \times 1$ convolution layer, respectively. To investigate the added value of TPM in the outcome prediction model, we trained and compared the performance of models with the following inputs:

1. CT and/or PET
2. CT and/or PET with manually segmented GTVp
3. CT and/or PET with tumor probability map (TPM)

Throughout the text, we will refer to single-modality when considering CT- or PET-only as input, and to multi-modality when multiple inputs are concatenated in the channel domain (all the other cases listed above).

The model outputs the risk score for each patient.

2.4.2. Optimization

A Cox negative logarithm partial likelihood loss [22] with L_2 regularization was used to optimize our model, which is defined as follows:

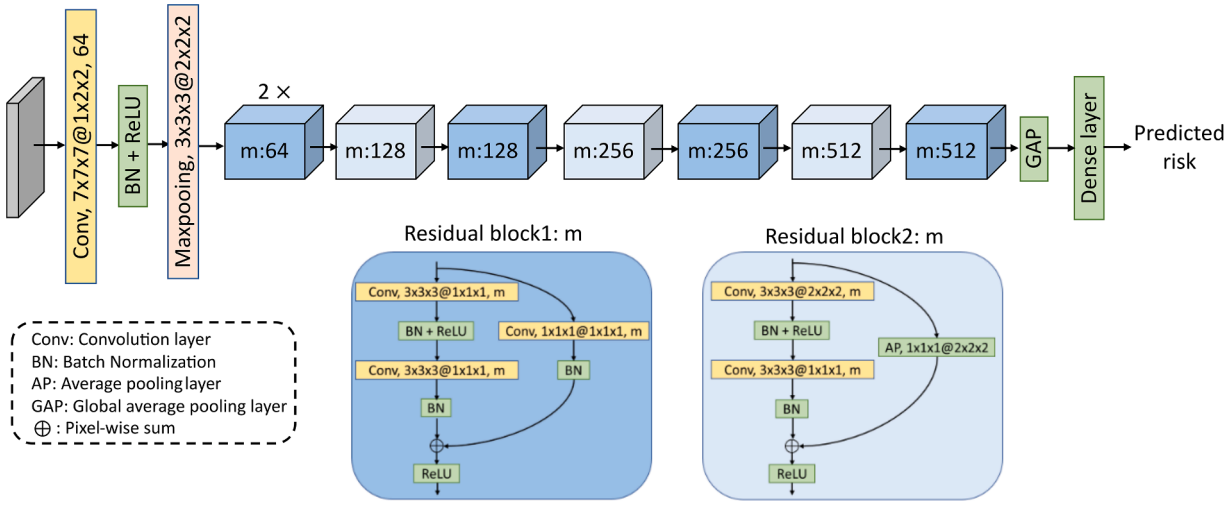

Fig. 2. The architecture of our outcome prediction model. The notation "$a \times b \times c$ @ $d \times e \times f$" denotes a convolutional kernel size of $a \times b \times c$ with a stride of $d \times e \times f$ along respective dimensions. m is the number of convolutional kernels used in residual blocks and C the number of channels. The multiplier at the top of every residual block indicates the number of times such block is repeated.

$$L = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(h_i - \log \sum_{j \in R(T_i)} e^{h_j} \right) + \lambda \| \theta \|_2 \quad (1)$$

where h is the predicted risk of one outcome endpoint, E is the event indicator (0 indicates a censored patient and 1 indicates a patient with event), T is the time-to-event (for $E = 1$) or time-to-censored (for $E = 0$). $N_{E=1}$ is the number of patients with event and $R(T_i)$ is a set of patients whose $T \geq T_i$, θ are parameters of the DL network and λ controls the contribution of L_2 regularization.

2.4.3. Experiment setup

The original training set was split into 3-folds stratified based on treatment date/year, to reduce systematic changes between the folds due to changes in imaging, treatment and other methods over time. Then, we used 3-fold cross-validation, resulting in three models whose predictions in the test set were average ensembled to obtain final test predictions. The models were trained using an SGD optimizer with a momentum of 0.99 for 400 epochs using PyTorch 1.6.0 and MONAI 0.8.1 packages on a Tesla V100 32 G GPU. The initial learning rate was 0.0002 which decreased by a factor of 0.2 in the 200th and 300th training epochs. To avoid over-fitting, early stopping with a patience number of 50 was the strategy to select the models with best validation C-index. Additionally, data augmentation was performed by random flip in three directions with a probability of 0.5 ($p = 0.5$), random affine transformation (rotation degree: 7.5, scale range [0.93–1.07], translation range: [0-7] and $p = 0.5$) and random elastic distortion (rotation degree: 7.5, scale range [0.93–1.07], translation range: 7 and $p = 0.2$). Oversampling was used to balance the number of censored patients and patients with events in the training.

To further improve outcome prediction performance, a hybrid Cox prediction model was constructed for each outcome endpoint using the predictions of the DL models and clinical Cox models based on clinical parameters only (Table A.1).

2.5. Model evaluation

To assess the quality of the tumor probability prediction models, the surface dice similarity coefficient with the GTVp contour at a tolerance of 3 mm was calculated for different probability thresholds. The C-index [95 % confidence interval (CI)] was used to assess the predictive performance of the models. The C-indexes obtained between different models were compared using the Z-test on 1000 bootstrapping samples. Calibration curves at 2-year after treatment were determined for the models achieving the highest C-index value in the test set. Then, the ability of the models to stratify patients into a high-risk (risk > threshold risk in the training set) and a low-risk (risk <= threshold risk) was evaluated by the log-rank test [23]. The threshold risk was determined by two steps: (1) calculating median risks for the patients' group with event and the patients' group without event, respectively and (2) obtaining the threshold risk by averaging these two median risks. Two-tailed p-value < 0.05 was considered significant.

3. Results

The demographics and patient characteristics of the training and test set used for the outcome prediction model are summarized in Table A.1. There were no significant differences between the two sets except for T-stage (T4 patients: 47.6 % vs. 37.0 %).

In Fig. 3 the DL network segmentation performance in the test sets are reported for different tumor probability thresholds. The overall trend suggests that the higher the threshold, the closer the output is to the GTVp manually delineated by the radiation oncologists. Using a probability threshold of 0.9, the median and the mean surface DSC are 0.88 and 0.74. At the same threshold, 72 % of all patients used in the outcome prediction model show a surface DSC result above the mean value. Cases which are outliers for a specific threshold are not always outliers for the other probability thresholds, caused by volumes with lower predicted probabilities in the GTVp area. Note that the subsequent DL network for outcome prediction uses the probability map as is, without using a cut-off threshold.

Table 1 shows the C-index values in the test set for the DL, clinical and best hybrid models. The latter are the combination of clinical and best DL models for each outcome endpoint prediction. The C-index values of all hybrid models using predictors of all the DL models are shown in Table A.2.

For all endpoints, the highest C-index results were obtained by hybrid and DL models. The **hybrid models** achieved the highest C-indexes of 0.61, 0.81 and 0.82 for RC, DMFS and OS, respectively. Only for LC the performance of the hybrid model (C-index: 0.71) was worse than that of the best **DL model** (C-index: 0.74). The **clinical models** for LC, RC and DFMS had lower performance than the best DL models. Only for OS the clinical model had a better performance than the best DL model with a C-index of 0.80.

To assess the effect of using TPM as additional input modality, a comparison with DL and hybrid models trained without TPM was made.

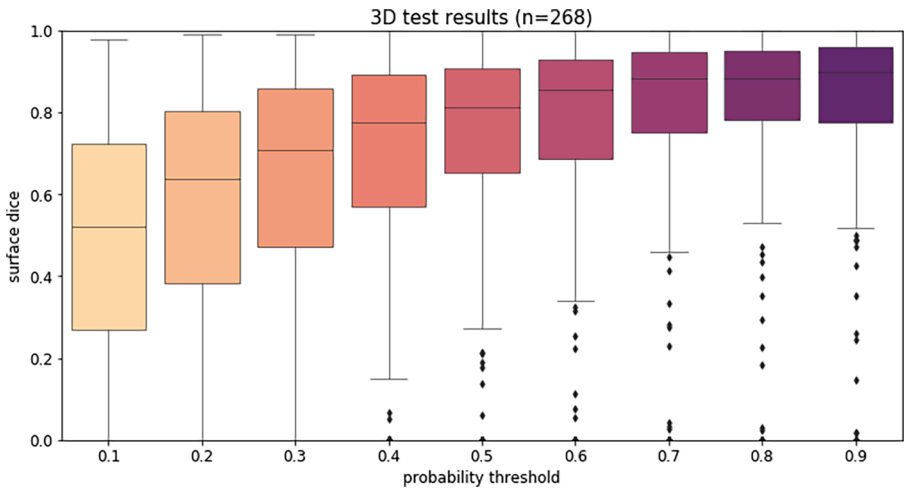

Fig. 3. Surface dice similarity coefficient with a tolerance of 3 mm calculated for different probability thresholds from the tumor probability map of patients used for the outcome prediction model.

Table 1
C-index values [95 % confidence interval (CI)] in the test set achieved by clinical, DL, and hybrid models for different input data. The hybrid models are obtained by the clinical models and the best DL models for each endpoint. In each column of the table, the highest C-index achieved by DL models is indicated in bold, and the highest C-index achieved by clinical, DL, and hybrid models is underlined.

Model	Input data	LC	RC	DMFS	OS
Clinical	T-stage, HPV status	0.64 [0.45,0.81]	–	–	–
	T-stage, N-stage	–	0.56 [0.44,0.69]	–	–
	N-stage	–	–	0.65 [0.59,0.74]	–
	T-stage, HPV status, WHO PS	–	–	–	0.80 [0.72,0.88]
Deep Learning	CT	0.62 [0.44,0.78]	0.51 [0.37,0.51]	0.60 [0.55,0.84]	0.71 [0.61,0.80]
	CT, GTVp	0.62 [0.41,0.82]	0.38 [0.25,0.53]	0.64 [0.46,0.81]	0.73 [0.62,0.83]
	CT, TPM	0.64 [0.45,0.81]	0.55 [0.42,0.68]	0.60 [0.44,0.76]	0.67 [0.56,0.77]
	PET	0.63 [0.45,0.78]	0.54 [0.41,0.67]	0.62 [0.47,0.78]	0.67 [0.56,0.78]
	PET, GTVp	0.66 [0.49,0.81]	0.55 [0.43,0.69]	0.59 [0.43,0.76]	**0.74 [0.64,0.84]** *
	PET, TPM	0.70 [0.52,0.86]	**0.60 [0.47,0.72]** *	0.52 [0.38,0.75]	0.72 [0.61,0.83]
	CT, PET	0.66 [0.46,0.83]	0.49 [0.34,0.63]	**0.73 [0.52,0.90]** *	0.59 [0.44,0.73]
	CT, PET, GTVp	0.69 [0.51,0.84]	0.53 [0.41,0.67]	0.47 [0.29,0.68]	0.63 [0.49,0.77]
	CT, PET, TPM	**0.74 [0.60,0.86]** *	0.56 [0.42,0.71]	0.63 [0.43,0.80]	0.68 [0.59,0.78]
Hybrid		0.71 [0.51,0.87]	0.61 [0.49,0.74]^	0.81 [0.65,0.92]^	0.82 [0.73,0.90]^

* denotes that this DL model has significant higher C-index than all other DL models. ^ denotes that this hybrid model has significant higher C-index than all other clinical and DL models.

The DL models using PET/CT with TPM and PET with TPM achieved the highest C-index of 0.74 and 0.60 **for LC and RC prediction**, respectively. For RC, the hybrid model obtained by the latter DL model and the clinical model resulted in a slightly higher C-index of 0.61. **For DMFS prediction**, the addition of GTVp or the TPM in the model training caused a decrease in C-index to 0.47 and 0.63, respectively. The DL model trained with PET and CT had the largest C-index of 0.73. **For OS prediction** the DL model trained with PET and GTVp had the highest C-index. However, models based on CT + GTVp, PET + TPM, and CT-only had similar performance with C-indexes of 0.73, 0.72, and 0.71, respectively. In Table A.2, it is shown that the hybrid models for OS based on PET + GTVp and clinical features had similar performance as the hybrid model based on PET + TPM and clinical features (C-index of 0.82 and no significant difference in Table A.2).

The calibration curves at 2-year after treatment for each endpoint are shown in Fig. 4. For 2-year OS, LC, and DMFS the values for the slope and intercept varied between 0.8 and 1.2, and between −0.2 and 0.2, respectively. For 2-year RC the calibration curve shows a slightly worse agreement between the predicted and observed probabilities compared to the other endpoints.

In Fig. 5 the Kaplan Meier curves for high and low risk patients stratified by the best DL and the hybrid models in the test set are drawn for LC, RC, DMFS, and OS. The KM curves for the clinical models are shown in Figure A.1. The clinical models could stratify patients into risk groups with significant difference in OS only. The DL models showed a good separation between the risk groups with significant differences in LC and OS, but no significant differences for DMFS. Whereas the hybrid models resulted in risk groups with significant differences for LC, DMFS, and OS.

4. Discussion and conclusions

We developed and compared deep learning based prediction models for LC, RC, DMFS and OS in OPSCC patients using different combinations of imaging modalities as input. We tested single- (PET or CT) and multi-modality (PET/CT) as well as their combination with the GTVp or the TPM obtained by training a DL model for tumor segmentation. Additionally, we built hybrid models that used both the predictions of the best DL models and clinical models to obtain more accurate outcome predictions.

The highest C-index results were achieved by hybrid models in OS, DMFS and RC prediction (0.61, 0.81 and 0.82) and by a DL model in LC prediction (0.74). Apparently clinical information about the T-stage and HPV status, which were the most significant clinical predictors for **LC**

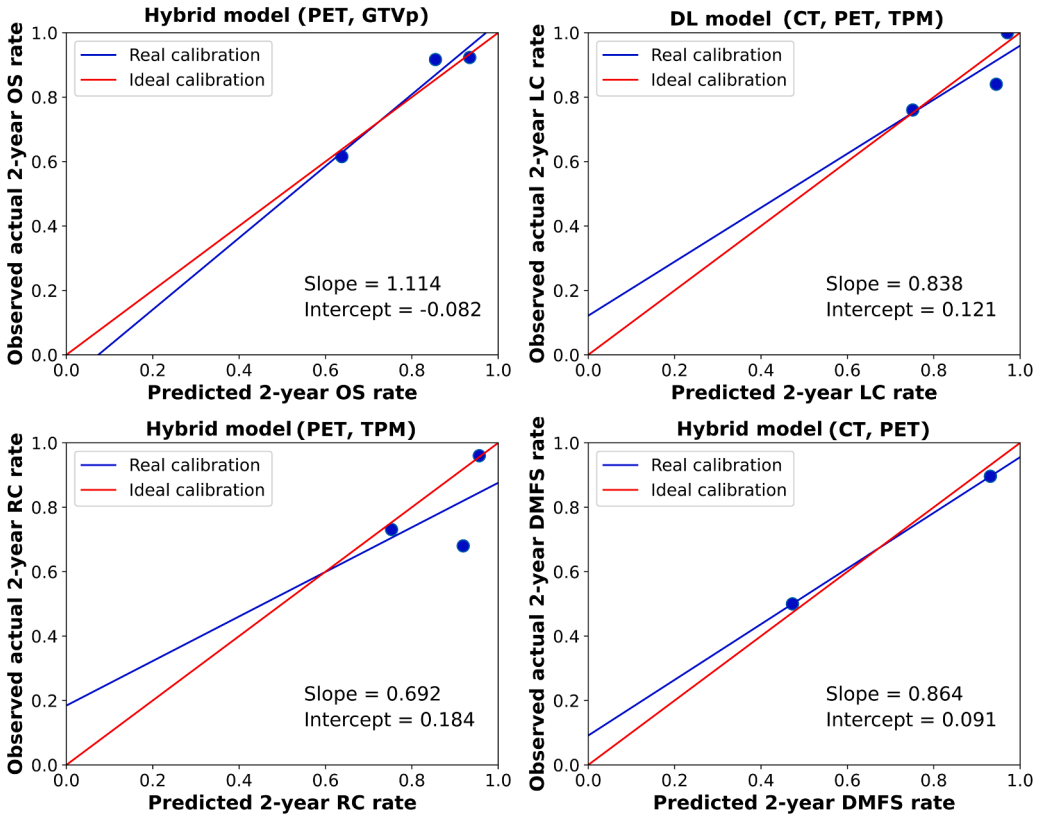

Fig. 4. Calibration curves at 2-year after treatment in the test sets for the best performing models from Table 1.

(Table A.1), did not add to the predictive performance of the tumor features extracted by the DL model. In the **OS** prediction the hybrid model based on the PET, GTVp DL model combined with the clinical features T-stage, HPV status and WHO PS had the highest C-index. However, a similar performance (C-index = 0.82 with no significant difference) was observed after replacing the GTVp with the TPM (Table 1). In the case of similar performance, the model based on TPM is preferred for a faster OS prediction because it does not need time-consuming manual tumor delineations. For **DMFS** prediction, the DL model using PET/CT as input achieved a higher C-index of 0.72 than other input combinations in the DL models. We hypothesized that DMFS prediction may rely mainly on lymph nodes features (only N-stage (not T-stage) is a significant predictor in the clinical model (Table A.1)). The GTVp does not contain lymph node structures and the TPM only partly and, in those cases, with very low probabilities. Thus, the addition of TPM or GTVp limits the network resulting in a decreased performance.

DL models that used **TPM as input,** in conjunction with single- or multi-modality images, achieved higher or comparable C-index values (0.74, 0.60 and 0.72 in Table 1) than DL models using GTVp (0.69, 0.55 and 0.74) for LC, RC and OS prediction. We attributed this to the nature of the TPM: areas corresponding to the GTVp have higher pixel values compared to the surrounding areas and/or in correspondence of lymph nodes with high metabolic activity on PET, where the probabilities are lower. These properties of the TPM could make the network focus on extracting features in regions with high probabilities of being tumor as well as considering features in low-probability regions. Using the binary segmentation mask of the GTVp, in conjunction with single- or multi-modality images, could direct the network in extracting features in the GTVp region only, neglecting useful information in the surrounding regions and not accounting for delineation uncertainty. Furthermore, the training of the DL model, could then be dominated by learning the GTVp shape features instead of CT and PET features, also shown by the tendency of the model to converge on the binary GTVp.

It was not the first time that a tumor probability map was used as input for an outcome prediction model. In the multi-task approach proposed by Meng et al. [14], the non-thresholded output of the segmentation backbone was used as additional input channel for the cascaded survival network. However, the nature of the probability map itself is one of the main differences between our study and the DeepMTS. Specifically, the tumor probability maps employed in our current study were obtained as in [20], widening the range of probabilities typically derived from a segmentation network's output. The network architecture and the operation of ensembling multiple predictions resulted, in fact, in probability maps that capture the segmentation model's uncertainty. This is the first study where the effect of tumor probability maps capturing the segmentation model's uncertainty is explored and compared to manually delineated structures in outcome prediction models.

In this study, we used single- and multi-modality as input of DL outcome prediction models. The rationale behind using multi-modal data is to provide the network with information from different sources. During the second edition of the HECKTOR challenge [10], participants were asked to build models for progression-free survival prediction using PET/CT images and available clinical data. Saeed et al. [24], who ranked first in the challenge, proved that the way multiple input data types are combined also plays a big role in the model performance. In their study they created a new image modality by averaging normalized PET and CT images and they achieved higher C-index compared to processing each modality separately and then combining the feature vectors in a later stage. In a more recent work by Wang et al. [25], multi-channel PET/CT DL models achieved worse performance compared to PET-only DL models in predicting DM and OS. In our work we combined multiple modalities concatenating them in the channel domain. We showed that for all endpoints the best DL models were always multi-modal (PET/CT or PET and/or CT combined with GTVp or TPM). However, CT-only and PET-only models did not always achieve

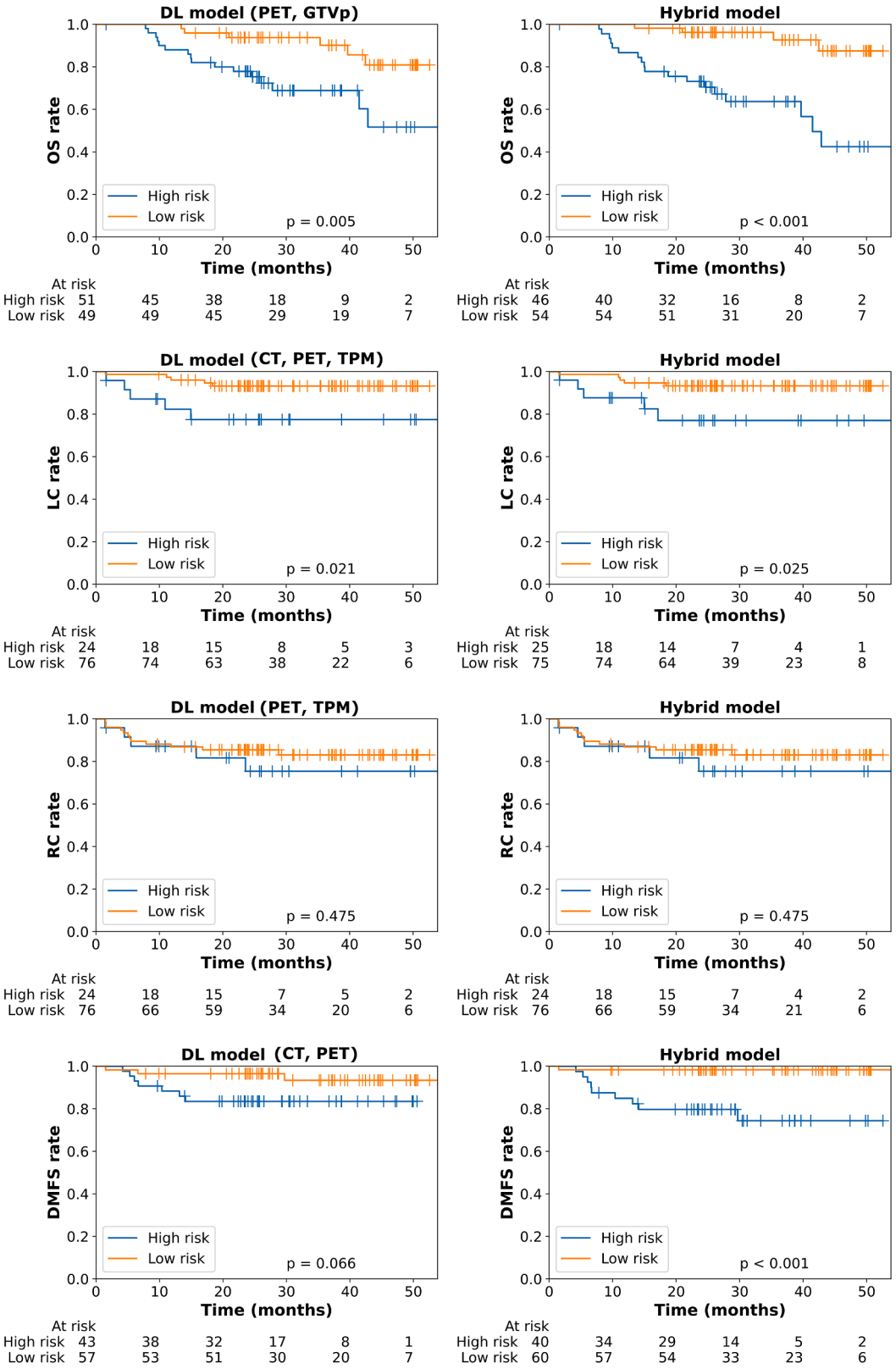

Fig. 5. KM-curves of risk stratification in the test sets achieved by best DL models and hybrid models.

the worst performance. It is still unclear whether adding more image modalities would represent a benefit or just redundant information for each of the endpoints. It is also possible that the way that images are concatenated in the input affects the model in extracting modality-specific features. To date, only a limited number of studies focused on outcome prediction in oropharyngeal cancer where different combinations of input data are compared, therefore further research is still needed to draw definite conclusions. To the best of our knowledge, this study represents the first comprehensive evaluation and comparison of the impact of utilizing various combinations of input data on the

prediction of diverse outcome endpoints in oropharyngeal cancer.

Our study supports other papers' hypothesis that the surrounding of the tumor contains prognostically relevant information [10,14,9,25]. In the paper from Starke et al. [26], they demonstrated that a model based on radiomic features calculated on a ROI obtained simply thresholding the Standardized Uptake Value (SUV) to 2.5 in the PET outperformed the expert tumor delineation for PFS. Our DL model for tumor probability prediction is highly dependent on PET SUV values, however it still provides joint PET/CT information. Feature maps created from DL models could help understand how these models extract features and support our hypothesis that TPM may provide more information than the GTVp for improving outcome prediction. Fig. 6 displays the input and feature maps of the DL models generated by averaging all feature maps output from the max pooling layer. Compared with GTVp (c), which just provides the tumor shape information delineated by a radiation oncologist, in TPM (d) each voxel represents a tumor probability which is generally consistent with PET (b) SUV values. Thus, we hypothesized that the TPM may guide DL training to focus on tumor regions with strong metabolic activity (high SUV values and high probabilities of being tumor) and to peri-tumor, low uptake tumor region, or false positives (low-probability of being tumor or lymph nodes). The PET only model (e) extracted image features mainly from high-active PET regions. When adding the GTVp the model tended to extract more GTV shape features than PET features because the GTVp is a simple binary mask which makes the model converge more easily. The shape of the feature map (g) showed a larger high intensity area compared to (e), which suggests that TPM helped the DL model extract features from both high-active tumor and peri-tumor regions. Fig. 6 (h) shows that the model trained on PET/CT did not focus on the entire tumor volume. When the GTVp was added (i), the model still relied both on the GTV shape information and irrelevant information from non-tumor/background regions. The combination of PET/CT and TPM (j) mainly highlights regions corresponding to the TPM, especially the ones with high probabilities of being tumors.

There are not many studies in literature where DL models are implemented for LR or RC prediction, especially using PET/CT as input. In Starke et al. [27] CT, GTVp, and clinical data were used as input of a 3D CNN for LRC prediction in HNC patients, achieving a C-index of 0.69.

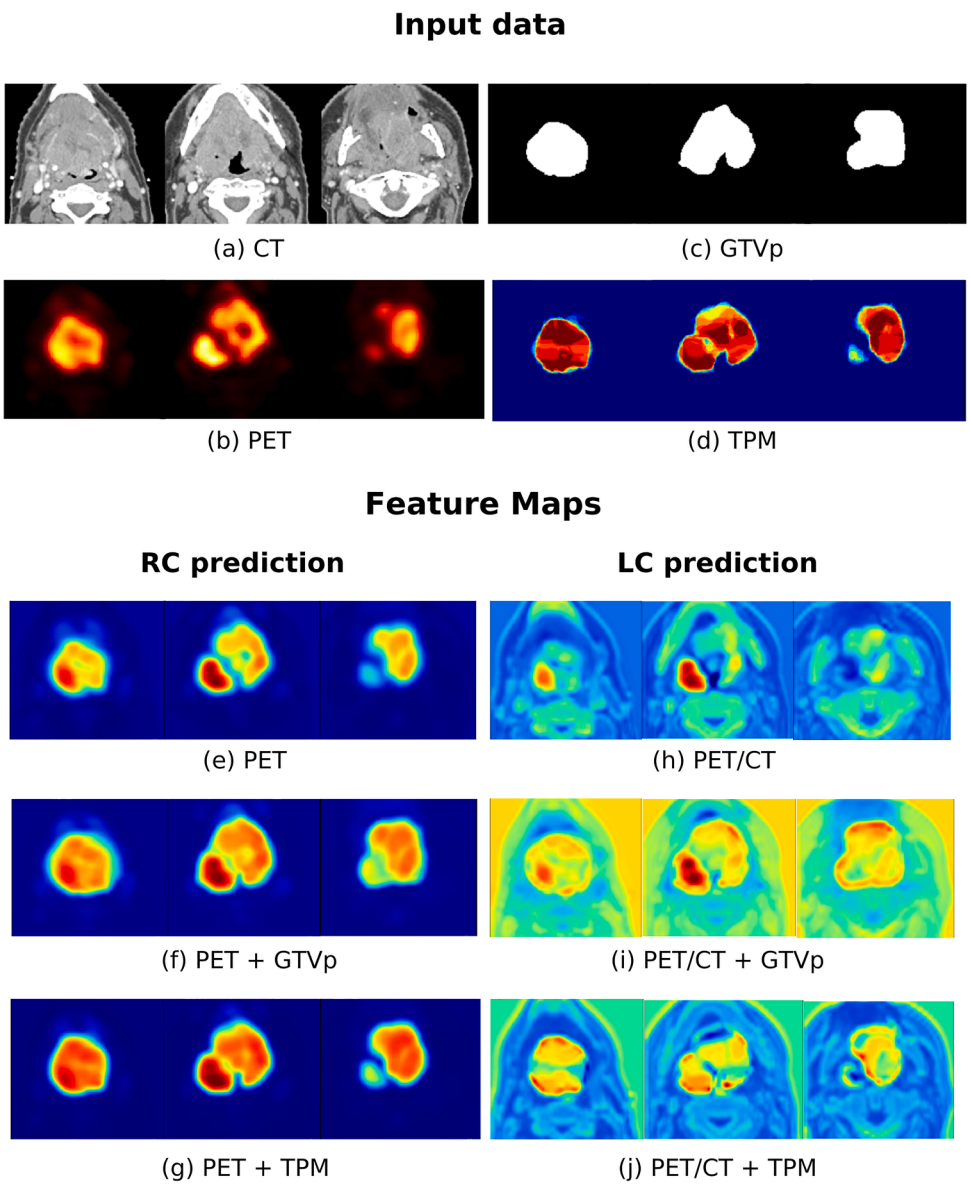

Fig. 6. Example of input image CT (a), PET (b), GTVp (c), TPM (d) and feature maps of DL models trained using as input PET (e), PET and GTVp (f), PET and TPM (g), PET/CT (h), PET/CT and GTVp (i) and PET/CT and TPM (j), respectively. The feature maps (e), (f), and (g) are from the DL model trained for RC prediction; (h), (g), and (j) are from the DL model trained for LC prediction.

In our experiments, the hybrid model which used CT and GTVp had comparable performance (C-index of 0.66). We showed that LC could be improved by inputting TPM together with the PET/CT, giving our best C-index of 0.74 for LC. For DMFS we obtained the best performance using CT and PET images in combination with clinical variables (C-index: 0.82), whereas Wang et al. obtained a similar model performance using PET only (C-index: 0.82) [25]. For OS, a previous study [3] proposed a DL-based fully automated GTV segmentation and prediction model in OPC patients using PET images and achieved a C-index of 0.79. Our best C-index of 0.82 is slightly higher and it was achieved by hybrid models trained on PET with GTVp and PET with TPM combined with clinical features.

We also investigated the risk stratification ability of clinical models, best DL models and hybrid models for LC, RC, DMFS and OS. No significant differences between high and low risk groups were observed for RC, which is reasonable because the highest achieved C-index value of 0.61 was too low. The clinical model only showed significant difference in OS prediction with a C-index of 0.80. Best DL models showed good risk stratifications ($p < 0.05$) for LC (PET/CT + TPM DL model) and OS (PET + GTVp DL model) with both high C-indexes of 0.74. Best hybrid models showed good risk stratification for LC, DMFS (based on PET/CT DL model) and OS. The above results showed our best DL and hybrid models could be applied to evaluate patient risk and select patients for individualized treatment. To summarize, TPM contributed to improve the risk stratification for LC only.

Some limitations exist in this study. Firstly, the DL model for tumor probability prediction failed in assigning high probabilities to GTVp pixels where no PET uptake is present (see Fig. A.2). This lack of consistency could have lowered the performance of the DL model for outcome prediction. Secondly, this study only focused on the primary tumor, so the image features from the lymph nodes, which are important for RC and DMFS prediction, were not considered. Additionally, for a few patients the bounding box was not big enough to include the entire primary tumor. Finally, all prediction models were built and tested on a small local UMCG data set of 268 patients. Multi-center data is expected to build a more robust model for external validation. In the future, the tumor probability prediction method could be improved, the lymph node structures could be included, and a better multi-modality feature extraction approach or adding MRI and dose images could be considered to improve the predictive performance of DL models.

In conclusion, we built deep learning-based prediction models using CT, PET, GTVp and the predicted tumor probability maps (TPM) for LC, RC, DMFS and OS prediction in OPSCC patients. Adding the TPM to image modalities as CT, PET or PET/CT demonstrated to improve the LC and RC prediction with a C-index of 0.74 and 0.61, respectively. For OS prediction, using the TPM or GTVp in combination with PET/CT and clinical data resulted in the same C-index of 0.82. Our models may serve as a potential tool for selecting patients for personalized treatment in order to improve treatment outcomes.

CRediT authorship contribution statement

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

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This research was supported by the Hanarth Fonds and by Chinese Scholarship Council (CSC). We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107939.

References

[1] K.L. Gage, K. Thomas, D. Jeong, D.G. Stallworth, J.A. Arrington, Multimodal imaging of head and neck squamous cell carcinoma, Cancer Control 24 (2) (2017) 172–179.
[2] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, J. Seuntjens, Deep learning in head & neck cancer outcome prediction, Sci. Rep. 9 (1) (2019).
[3] N.M. Cheng, et al., Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using FDG-PET imaging, Clin. Cancer Res. 27 (14) (2021) 3948–3959.
[4] E. Lombardo, et al., Distant metastasis time to event analysis with CNNs in independent head and neck cancer cohorts, Sci. Rep. 11 (1) (2021).
[5] H. Zheng, et al., Multi-transSP: multimodal transformer for survival prediction of nasopharyngeal carcinoma patients, Lect. Notes Comput. Sci. 13437 (2022) 234–243.
[6] N. Saeed, I. Sobirov, R. Al Majzoub, M. Yaqub, TMSS: an end-to-end transformer-based multimodal network for segmentation and survival prediction, Lect. Notes Comput. Sci. 13437 (2022) 319–329.
[7] X.D. Huang, et al., Competing risk nomograms for nasopharyngeal carcinoma in the intensity-modulated radiotherapy era: a big-data, intelligence platform-based analysis, Radiother. Oncol. 129 (2) (2018) 389–395.
[8] T. Valencia, et al., Metabolic reprogramming of stromal fibroblasts through p62-mTORC1 signaling promotes inflammation and tumorigenesis, Cancer Cell 26 (1) (2014) 121.
[9] X. Wang, et al., Can peritumoral regions increase the efficiency of machine-learning prediction of pathological invasiveness in lung adenocarcinoma manifesting as ground-glass nodules? J. Thorac. Dis. 13 (3) (2021) 1327–1337.
[10] V. Andrearczyk, et al., Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images, Lect. Notes Comput. Sci. 13209 (2022) 1–37, https://doi.org/10.1007/978-3-031-27420-6_1.
[11] K.A. Wahid, 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, Lect. Notes Comput. Sci. 13209 (2022) 300–307, https://doi.org/10.1007/978-3-030-98253-9_28.
[12] P. Afshar, A. Mohammadi, K.N. Plataniotis, A. Oikonomou, H. Benali, From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities, IEEE Signal Process. Mag. 36 (4) (2019) 132–160.
[13] V. Andrearczyk, et al., Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer, Lect. Notes Comput. Sci. 12928 (2021) 147–156.
[14] M. Meng, B. Gu, L. Bi, S. Song, D.D. Feng, J. Kim, DeepMTS: deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT, IEEE J. Biomed. Health Inform. 26 (9) (2022) 4497–4507.
[15] M. Meng, Y. Peng, L. Bi, J. Kim, Multi-task deep learning for joint tumor segmentation and outcome prediction in head and neck cancer, Lect. Notes Comput. Sci. 13209 (2022) 160–167.
[16] M. Meng, L. Bi, D. Feng, J. Kim, Radiomics-enhanced deep multi-task learning for outcome prediction in head and neck cancer, Lect. Notes Comput. Sci. 13626 (2023) 135–143.
[17] J. Ren, B.N. Huynh, A.R. Groendahl, O. Tomic, C.M. Futsaether, S.S. Korreman, PET normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT, Lect. Notes Comput. Sci. 13209 (2022) 83–91.
[18] A. De Biase, N.M. Sijtsema, L. van Dijk, J.A. Langendijk, P. van Ooijen, Slice-by-slice deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for spatial uncertainty on FDG PET and CT images, arXiv (2022), https://doi.org/10.48550/ARXIV.2207.01623.
[19] L.V. van Dijk, et al., 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia, Radiother. Oncol. 126 (1) (2018) 89–95.
[20] A. De Biase, N.M. Sijtsema, L.V. van Dijk, J.A. Langendijk, P.M.A. van Ooijen, Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for

predicted tumor probability in FDG PET and CT images, Phys. Med. Biol. 68 (5) (2023), 055013.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.
[22] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (1) (2018).
[23] N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration, Cancer Chemother. Rep. 50 (3) (1966).
[24] N. Saeed, R. Al Majzoub, I. Sobirov, M. Yaqub, An ensemble approach for patient prognosis of head and neck tumor using multimodal data, Lect. Notes Comput. Sci. 13209 (2022) 278–286.
[25] Y. Wang, et al., Deep learning based time-to-event analysis with PET, CT and joint PET/CT for head and neck cancer prognosis, Comput. Methods Programs Biomed. 222 (2022).
[26] S. Starke, D. Thalmeier, P. Steinbach, M. Piraud, A hybrid radiomics approach to modeling progression-free survival in head and neck cancers, Lect. Notes Comput. Sci. 13209 (2022) 266–277.
[27] S. Starke, et al., 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma, Sci. Rep. 10 (1) (2020).