SCIENTIFIC REPORTS

natureresearch

OPEN

2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma

Sebastian Starke[1,2,3]✉, Stefan Leger[2,3,4], Alex Zwanenburg[2,3,4], Karoline Leger[2,3,4,5], Fabian Lohaus[2,3,4,5], Annett Linge[2,3,4,5], Andreas Schreiber[6], Goda Kalinauskaite[7,8], Inge Tinhofer[7,8], Nika Guberina[9,10], Maja Guberina[9,10], Panagiotis Balermpas[11,12], Jens von der Grün[11,12], Ute Ganswindt[13,14,15,16], Claus Belka[13,14,15], Jan C. Peeken[13,17,18], Stephanie E. Combs[13,17,18], Simon Boeke[19,20], Daniel Zips[19,20], Christian Richter[2,3,5,21], Esther G. C. Troost[2,3,4,5,21], Mechthild Krause[2,3,4,5,21], Michael Baumann[2,3,4,5,21,22] & Steffen Löck[2,3,5]

For treatment individualisation of patients with locally advanced head and neck squamous cell carcinoma (HNSCC) treated with primary radiochemotherapy, we explored the capabilities of different deep learning approaches for predicting loco-regional tumour control (LRC) from treatment-planning computed tomography images. Based on multicentre cohorts for exploration (206 patients) and independent validation (85 patients), multiple deep learning strategies including training of 3D- and 2D-convolutional neural networks (CNN) from scratch, transfer learning and extraction of deep autoencoder features were assessed and compared to a clinical model. Analyses were based on Cox proportional hazards regression and model performances were assessed by the concordance index (C-index) and the model's ability to stratify patients based on predicted hazards of LRC. Among all models, an ensemble of 3D-CNNs achieved the best performance (C-index 0.31) with a significant

[1]Helmholtz-Zentrum Dresden - Rossendorf, Department of Information Services and Computing, Dresden, Germany. [2]OncoRay - National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany. [3]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Dresden, Dresden, Germany. [4]National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany. [5]Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany. [6]Department of Radiotherapy, Hospital Dresden-Friedrichstadt, Dresden, Germany. [7]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Berlin, Berlin, Germany. [8]Department of Radiooncology and Radiotherapy, Charité University Hospital, Berlin, Germany. [9]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Essen, Essen, Germany. [10]Department of Radiotherapy, Medical Faculty, University of Duisburg-Essen, Essen, Germany. [11]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Frankfurt, Frankfurt, Germany. [12]Department of Radiotherapy and Oncology, Goethe-University Frankfurt, Frankfurt, Germany. [13]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Munich, Munich, Germany. [14]Department of Radiation Oncology, Ludwig-Maximilians-Universität, Munich, Germany. [15]Clinical Cooperation Group, Personalized Radiotherapy in Head and Neck Cancer, Helmholtz Zentrum, Munich, Germany. [16]Department of Radiation Oncology, Medical University of Innsbruck, Anichstraße 35, 6020 Innsbruck, Austria. [17]Department of Radiation Oncology, Technische Universität München, Munich, Germany. [18]Institute of Radiation Medicine (IRM), Helmholtz Zentrum München, Neuherberg, Germany. [19]German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Tübingen, Tübingen, Germany. [20]Department of Radiation Oncology, Faculty of Medicine and University Hospital Tübingen, Eberhard Karls Universität Tübingen, Tübingen, Germany. [21]Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiooncology – OncoRay, Dresden, Germany. [22]German Cancer Research Center (DKFZ), Heidelberg, Germany. ✉email: s.starke@hzdr.de

association to LRC on the independent validation cohort. It performed better than the clinical model including the tumour volume (C-index 0.39). Significant differences in LRC were observed between patient groups at low or high risk of tumour recurrence as predicted by the model ($p = 0.001$). This 3D-CNN ensemble will be further evaluated in a currently ongoing prospective validation study once follow-up is complete.

Treatment individualisation is a central objective for the improvement of radiotherapy outcomes[1]. In particular, patients diagnosed with locally advanced head and neck squamous cell carcinoma (HNSCC) might benefit from individualised treatment, since five-year overall survival probability after primary radiochemotherapy is only approx. 50%[2]. Subgroups of patients may be identified that are currently under- or overtreated and might benefit from e.g. escalated or de-escalated dose prescriptions. Individualisation of treatment may be based on statistical survival models that predict endpoints such as overall survival or loco-regional tumour control (LRC). Survival models are able to analyse time-to-event data which frequently contain censored observations. The prognostic value of these models is based on biomarkers that are able to stratify patients into groups at different risk of treatment failure. Such biomarkers may result from clinical or tumour-related features such as age, gender or tumour stage, molecular analyses of tumour biopsies such as human papillomavirus (HPV) status or gene signatures, dosimetric information or clinical imaging data from computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) scans or combinations thereof[3–13].

Imaging data are considered a valuable source of information for tailoring individual treatment, due to their non-invasiveness, repeatability and their ability to represent the entire tumour. Numerous radiomics models, in which traditional machine-learning (ML) methods were applied on hundreds to thousands of pre-defined and handcrafted image features, have been developed[14–18], but have not yet surpassed the threshold for clinical acceptance and applicability[19]. Recently, Ger et al.[20] found that radiomics features of CT and PET scans failed to improve upon clinical risk models in a large head and neck cancer dataset. With the recent advances that deep convolutional neural networks (CNNs) have brought to the fields of natural and medical image analysis, there is hope to elevate model performance for radiotherapy outcome modelling, as well. This is mostly due to the fact that CNNs are able to automatically learn abstract feature representations of the input data during training. However, so far most applications of deep learning to medical images revolve around tasks for segmentation[21] or classification[22–24]. The same holds true for the field of radiotherapy, where most applications of deep learning focus on segmentation, computer-aided detection or motion management[25]. Only few attempts have been published to combine deep learning on medical imaging data and survival analysis[26–28].

The Cox proportional hazards model (CPHM) is a clinically established survival model. It is often used because it allows to exclusively model effects of patient covariates on individual event times without making distributional assumptions. Previously, Katzman et al.[29] demonstrated the benefit of combining multi-layer perceptrons with the CPHM for modelling of nonlinear feature interactions, while Ching et al.[30] applied this approach to ten cancer-related datasets of high throughput transcriptomics data. Additionally, the CPHM and CNNs were combined to build risk models based on pathological histology images for lung cancer and glioblastoma patients, respectively[26,27], whereas Haarburger et al.[28] applied a similar idea to CT scans of lung cancer patients.

In this manuscript, we developed and independently validated three deep learning approaches to predict LRC from treatment-planning CT images of patients with locally advanced HNSCC treated by primary radiochemotherapy. (i) We developed a CPHM based only on clinical parameters to provide a baseline model. Subsequently, we investigated different deep learning approaches to the CPHM: (ii) we trained 3D- and 2D-CNNs from scratch, (iii) we applied a transfer learning strategy by fine-tuning pre-trained networks on our dataset and (iv) we used deep features[31–33] generated by a trained autoencoder[34].

Methods

Patient cohort.

A multicentre retrospective cohort consisting of 291 patients with locally advanced HNSCC was collected and divided into an exploratory and an independent validation cohort (206 and 85 patients, respectively). Allocation of patients was based on the different included studies. 149 of the 206 patients of the exploratory cohort were treated in one of the six partner sites of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG) between 2005 and 2011[7]. The remaining 57 patients were treated at the University Hospital Dresden (UKD, Germany) between 1999 and 2006[35]. 51 of the 85 patients of the independent validation cohort were treated within a prospective clinical trial (NCT00180180) at the UKD between 2006 and 2012[3, 9]. 20 additional patients were treated at the UKD and the Radiotherapy Center Dresden-Friedrichstadt between 2005 and 2009 and the remaining 14 patients were treated in Tübingen between 2008 and 2013[36].

All patients received a CT scan for treatment-planning and were treated by primary radiochemotherapy. Inclusion criteria have previously been described[3,7,9,35]. Ethical approval for the multicentre retrospective analyses was obtained from the Ethics Committee at the Technische Universität Dresden, Germany (EK177042017)[17]. All analyses were carried out in accordance with the relevant guidelines and regulations. Informed consent was obtained from all patients. CT scans were provided as DICOM files with contours of the primary tumour manually delineated and reviewed by experienced radiation oncologists (F.L., K.L., E.G.C.T.). Patient characteristics are summarised in Table 1.

The primary endpoint of this study was LRC, which was defined as the time between the start of radiochemotherapy and local or regional tumour recurrence. For patients with observed loco-regional recurrence, the event time was accompanied by an event indicator variable of 1, whereas for patients without an observed event, the last follow-up time was used together with an event indicator variable of 0.

	Exploratory cohort (n = 206)		Independent validation cohort (n = 85)		
Variable	Median	(Range)	Median	(Range)	*p* value
Follow up time of patients alive (months)	52.62	(4.27–131.91)	42.55	(7.85–107.27)	0.72
Age (years)	59.00	(39.20–84.50)	55.00	(37.00–76.00)	0.023
Primary tumour volume (cm^3)	29.13	(4.52–321.74)	40.62	(2.70–239.07)	0.039
	Number of patients	(%)	Number of patients	(%)	
Gender male/female	174/32	(84/16)	74/11	(87/13)	0.70
cT-stage T1/T2/T3/T4	2/23/51/130	(1/11/25/63)	2/9/30/44	(2/11/35/52)	0.21
cN-stage N0/N1/N2/N3/unknown	30/7/154/15/0	(15/3/75/7/0)	9/8/64/3/1	(11/9/75/4/1)	0.097
UICC-stage I/II/III/IV	0/0/15/191	(0/0/7/93)	1/2/9/73	(1/2/11/86)	0.039
Tumour site oropharynx/oral cavity/hypopharynx/larynx	93/51/62/0	(45/25/30/0)	29/23/28/5	(34/27/33/6)	0.003
p16 status negative/positive/unknown	148/28/30	(72/13/15)	52/5/28	(61/6/33)	0.26
Pathological grading 0/1/2/3/unknown	1/6/131/61/7	(1/3/63/30/3)	0/0/43/35/7	(0/0/51/41/8)	0.071
Smoking status no/yes/unknown	41/163/2	(20/79/1)	13/51/21	(15/60/25)	1.00
Alcohol consumption no/yes/unknown	62/85/59	(30/41/29)	23/25/37	(27/29/44)	0.60
Loco-regional tumour recurrence	84	(41)	28	(33)	0.26

Table 1. Patient characteristics of the exploratory and independent validation cohort: *p* values were obtained by using two-sided Mann–Whitney U-tests for continuous variables and χ^2 homogeneity tests for categorical variables.

Design of the analysis. In our analyses we applied the CPHM, which is a regression model commonly used in survival analysis for assessing endpoints like overall survival, progression-free survival or, as in our case, LRC. It is able to take into account heterogeneous event times and censored observations and hence does not require the specification of a predefined and fixed follow-up time. The CPHM assigns a hazard, i.e. a risk, to every patient for developing a loco-regional recurrence, which can subsequently be used to classify patients into different risk groups for loco-regional failure. The study design is presented in Fig. 1. We investigated four approaches to develop survival models based on the CPHM for the prediction of LRC hazards for patients diagnosed with locally-advanced HNSCC. First, (i) a clinical model was trained on the exploratory cohort and evaluated on the independent validation cohort to provide baseline performance metrics. Moreover, three deep learning based strategies using CNNs were applied: We (ii) trained models completely from scratch, using 3D-CNNs as well as 2D-CNNs, applied (iii) a transfer learning approach leveraging weights of pre-trained 2D-CNN networks, and created (iv) a deep autoencoder and used its bottleneck features in a traditional CPHM.

Prognostic performance was evaluated by two approaches, calculation of the concordance index (C-index) and the ability to stratify patients into two risk groups based on the model predictions. The C-index[37–39] measures the alignment between the observed times of loco-regional recurrence and the model predictions. It is given on a scale between zero and one with 0.5 indicating no prognostic value of the model. A C-index close to zero represents perfect predictions, since predicted hazards should be lower for patients with a longer recurrence-free time. We emphasise that this is in contrast to the situation of directly predicting event times, where a C-index close to one would be desirable. 95% confidence intervals (CI) for C-indices were computed using the survcomp R package[40,41] which implements the method proposed by Pencina et al.[42]. Models that did not contain the C-index 0.5 within the 95% CI on the independent validation cohort were considered as successfully validated.

Furthermore, based on the model predictions, patients were assigned to two groups, at low or at high risk for loco-regional recurrence. This stratification was based on the hazard values predicted by the models for every individual patient. The median value of these predictions on the exploratory cohort was used as a cutoff. Patients with a predicted hazard exceeding the cutoff were assigned to the high risk group and the remaining patients with hazards smaller or equal to the cutoff were assigned to the low risk group. To stratify patients of the independent validation cohort, the same cutoff was applied. The difference in LRC between the stratified patient groups was assessed using the log-rank test for the Kaplan–Meier (KM) curves of both risk groups. Significance was established for *p* values below 0.05.

To address the random nature of the CNN training procedure and to leverage the benefits of model ensembles[43], we repeated model training three times, each time using 10-fold cross-validation (CV) based on the exploratory cohort, stratified by the LRC event status, for a total of 30 CV runs. By applying CV on the exploratory cohort, splits of the samples into training and internal test folds were obtained. Models were built in each CV run using the data of the training fold. Data of the internal test fold was set aside for optional hyperparameter tuning and data of the independent validation cohort was used to measure model performance on previously unseen data.

Since each of the 30 CV runs resulted in a trained model (which we refer to as single model), we created ensemble predictions by averaging of the network outputs, essentially considering the information of multiple models before making a final prediction.

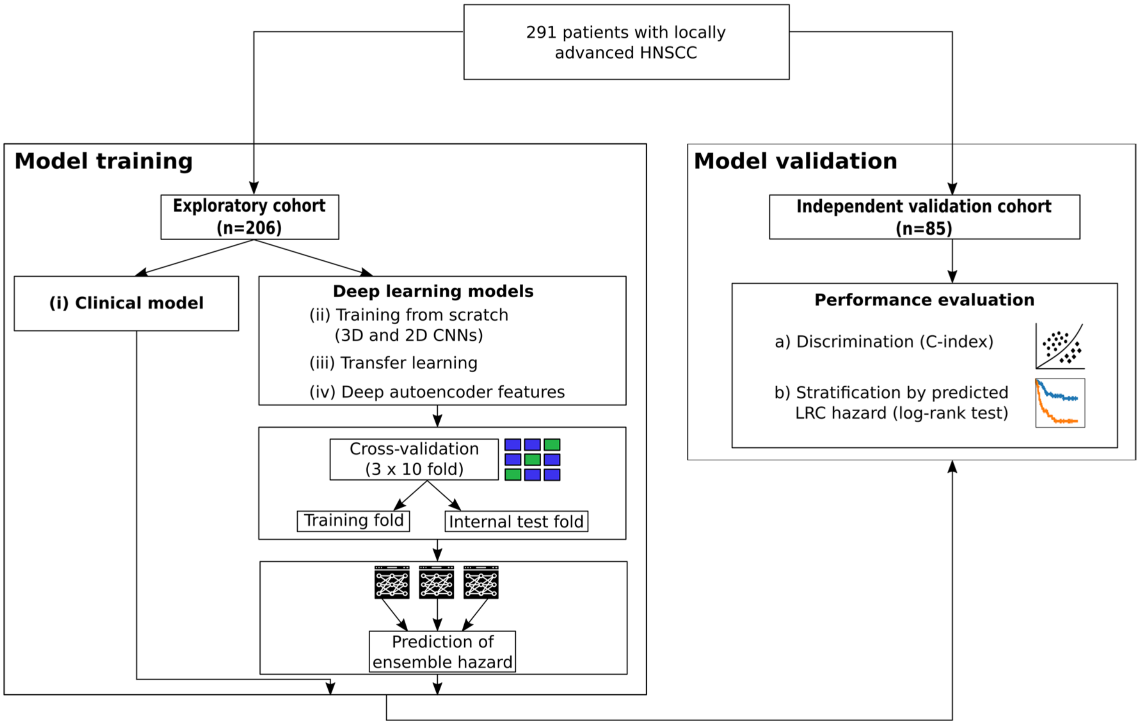
291 patients with locally advanced HNSCC
Model training
Exploratory cohort (n=206)
(i) Clinical model
Deep learning models
(ii) Training from scratch (3D and 2D CNNs)
(iii) Transfer learning
(iv) Deep autoencoder features
Cross-validation (3 x 10 fold)
Training fold
Internal test fold
Prediction of ensemble hazard
Model validation
Independent validation cohort (n=85)
Performance evaluation
a) Discrimination (C-index)
b) Stratification by predicted LRC hazard (log-rank test)

Figure 1. Design of the analysis. (i) To provide baseline results, a clinical Cox proportional hazards model (CPHM) was trained on the exploratory cohort and evaluated on the independent validation cohort. (ii)–(iv) Three deep learning approaches were evaluated by training convolutional neural networks in a cross-validation approach. Subsequently, for each approach ensembles were constructed from the models obtained during cross-validation and their performance was evaluated on the independent validation cohort.

Image processing.

Preprocessing of patient CT scans was carried out using an in-house developed toolkit[44] (available from https://github.com/oncoray/mirp) by performing (1.) cubic interpolation to isotropic voxel size of 1 mm^3, (2.) cropping of the transversal plane to 224 by 224 pixels (with the tumour's centre of mass as the centre of the cropped slice), (3.) clipping of the intensity range of Hounsfield units (HU) to the range [-200, 200] and (4.) normalisation of pixel values to the interval (0, 1).

Multiple image samples of each patient's CT scan were extracted and used for model training and prediction. For all 2D-CNN models, we used 7 slices cranial and 8 slices caudal of the slice with the largest tumour area as provided by the segmentation mask, comprising a total of 16 transversal CT slices per patient. For training of the 3D-CNNs we used smaller image regions of the axial plane due to GPU memory limitations. We first extracted a 32 × 64 × 64 (z × y × x) sized volume centered at the tumour centre of mass. Then, 15 additional random volumes of the same size were extracted for each patient. The volume centres were uniformly sampled from a cubic region of edgelength 32 around the tumour centre of mass. Zero padding was added to all extracted volumes where necessary. For each of the volumes, a prediction was computed. Those were subsequently averaged to obtain a single prediction for each patient.

Cox proportional hazards model.

The traditional CPHM fits the effect of p-dimensional covariates $\boldsymbol{x}$ on the hazard function h via $h(t,\boldsymbol{x}) = h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j\right)$, with an unspecified baseline hazard function $h_0(t)$. We followed Katzman et al.[29] in extending this to the more general form of $h(t,\boldsymbol{x}) = h_0(t)\exp\left(\gamma_{\boldsymbol{\beta}}(\boldsymbol{x})\right)$ with $\boldsymbol{\beta}$ denoting weights learned by a neural network. Log-hazard values $\gamma_{\boldsymbol{\beta}}(\boldsymbol{x})$ were estimated from CT image samples $\boldsymbol{x}$ by minimisation of (a batch approximation of) the negative of the Cox partial log-likelihood function

$$\ln L = \sum_{i=1}^{n} \delta_i \left(\gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - \ln\left(\sum_{\substack{j=1 \\ t_j \geq t_i}}^{n} \exp(\gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_j)) \right) \right), \tag{1}$$

letting δ_i denote an event indicator variable that takes on the value 1 if loco-regional tumour recurrence was observed for CT sample i and 0 otherwise, and n being the total number of available CT samples. Further details on survival analysis and the CPHM are given in "Survival analysis and deep Cox proportional hazards modelling" section of the supplement.

All computations were done using Python 3.6.7 and Keras 2.2.4[45] with tensorflow (v1.12.0) backend. Our code is publically available from https://github.com/oncoray/cnn-hnscc and experimental outputs can be downloaded from http://doi.org/10.14278/rodare.255.

Clinical model. To develop the clinical CPHM, we considered the clinical features patient age, gender, cT-stage, cN-stage, UICC-stage, tumour site, p16 status, pathological grading, smoking status, alcohol consumption and primary tumour volume. These features have already been considered in previous studies[7,35]. Tumour site comprised the values oropharynx, hypopharynx, larynx and oral cavity and was one-hot encoded. Volume was computed by summation of tumour segmentation masks and division by a factor of 1000 to obtain units of cm^3, followed by a (natural) logarithmic transformation.

Imputation of missing values for cN-stage, pathological grading and smoking status (1, 14 and 23 cases, respectively) was performed through selection of the most frequent value in the exploratory cohort. Due to more missing values (58 cases), p16 was converted into the variables $p16_{\text{unknown}}$ and p16. The same was done for alcohol consumption for which there were 96 missing cases. cT, cN, UICC and pathological grading stages were converted into the binary categories $cT < 4$, $cN < 2$, UICC < 4[7] and pathological grading < 2. Patient age and tumour volume were z-score normalised with means and standard deviations obtained from the exploratory cohort. Clinical features prognostic for LRC were selected by applying a forward variable selection CPHM based on the likelihood ratio test (inclusion $\alpha = 0.05$, exclusion $\alpha = 0.1$) using the exploratory cohort. Finally, a CPHM was trained on the exploratory cohort using the selected features and applied to the independent validation cohort.

Model ensembles. Due to our cross-validation approach (10-fold CV repeated three times), 30 different models were trained in every analysis. By averaging the resulting predicted log-hazard values, one final ensemble prediction for the hazard of loco-regional recurrence was obtained for every patient. On the independent validation cohort, a patient's ensemble prediction was computed by averaging over all 30 model predictions. For every patient of the exploratory cohort, a training and an internal test ensemble prediction was computed, since they appeared as part of the training folds and as part of the internal test folds. Training ensemble predictions were obtained by computing for every patient an average over all those 27 models for which that patient was part of the training fold. Similarly, internal test ensemble predictions were computed by only using the remaining three models for which the patient belonged to the internal test fold. For ensemble stratification of patients into groups at low and high risk of loco-regional recurrence, the cutoff value was determined as the median value of the training ensemble predictions.

Training from scratch. Different network architectures of 3D-CNN and 2D-CNN models were trained from scratch. In all trainings we used the AMSGrad version[46] of the Adam optimiser to estimate model parameters. For the 3D-CNN experiments the same architecture and hyperparameters as given by Hosny et al.[23] were used with small changes. Due to a different input shape, the first dense layer contained slightly fewer neurons. In the last layer, a single output neuron with tanh activation was used instead of two neurons with softmax activation which they used for classification purposes. Each model was trained for a fixed number of 200 epochs with a batch size of 24. Neither data augmentation nor callbacks for early stopping or learning rate adjustments were used.

The 2D-CNN architecture (Fig. 2) was loosely inspired by the VGG architecture[47]. It consisted of five convolution blocks, each containing two convolutional layers with filter size 5×5 for the first block and 3×3 for all remaining blocks and ReLU activation functions. No batch normalisation (BN) was used. The second convolutional layer of each block performed downsampling by using a stride of two. The first block comprised of 16 filters. The number of filters was doubled in each subsequent block. A flattening operation and a dropout layer (p=0.3) followed the last convolutional layer, before being connected to two fully-connected dense layers with ReLU activation of sizes 256 and 64, respectively. Dropout with the same probability as above was also applied between those dense layers. Lastly, the model output was given by a single dense neuron with tanh activation. Training was done for a maximum of 50 epochs with a learning rate of $5 \cdot 10^{-5}$ while doing early stopping (patience=10) on the internal test fold as well as reducing learning rates on plateaus (factor=0.5, patience=3, $min\_lr = 10^{-7}$) via the provided Keras callbacks. We also evaluated performance after replacing the final tanh activation with a linear output, essentially allowing for unrestricted log-hazard ranges. Moreover, the effect of inserting BN layers between convolutions and ReLU activations was assessed.

The effect of combining clinical features and CT samples as two separate inputs to a 2D-CNN was evaluated. First, Spearman correlation coefficients between the 2D-CNN model-output (with BN and tanh as final activation) and the clinical features were computed. Then, a second input branch, designed to estimate log-hazard values from the clinical features was added to the network architecture as depicted in Fig. 2b). It consisted of a single dense neuron with tanh activation and with BN. The log-hazard estimates coming from the clinical branch and the image branch were then concatenated and fed through the final output layer consisting again of a single dense neuron with tanh activation and with BN.

Transfer learning. We evaluated the capabilities of transfer learning for training 2D-CNNs. The ResNet50[48], DenseNet201[49] and InceptionResNetV2 (IRNV2)[50] architectures with weights pre-trained on the ImageNet dataset were used as foundation models. Their fully connected layers were replaced by a global average pooling layer followed by three dense layers with 128, 32 and one neurons, respectively. The first two dense layers utilised the ReLU activation function and the final layer used the tanh activation to restrict hazard output to the range $(\exp(-1), \exp(1))$. No BN was applied in the newly added layers.

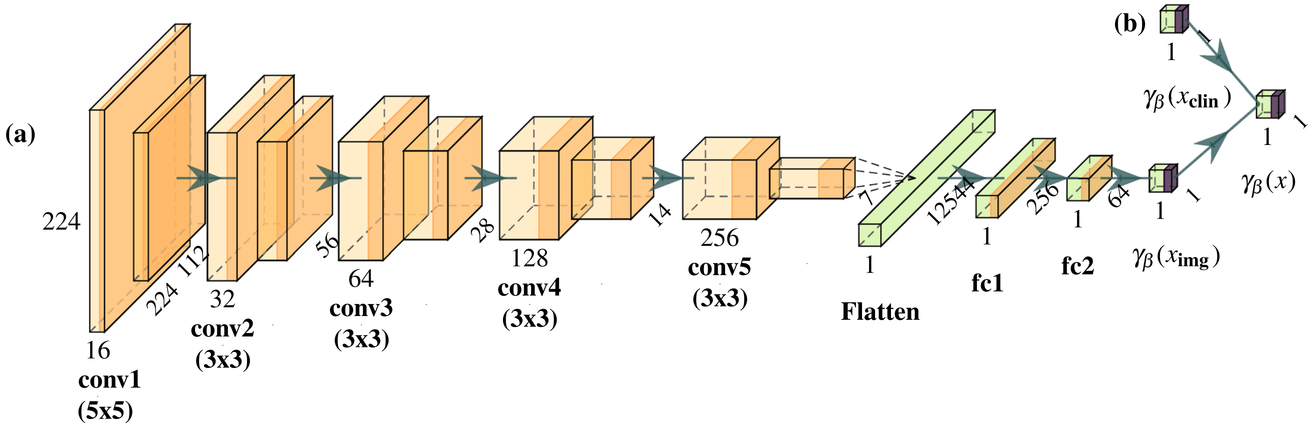

Figure 2. Architecture used when training a 2D-convolutional neural network from scratch. Numbers give shapes of computed feature maps. The network consists of convolutional filters ('conv', light orange), with ReLU activation functions (orange). These are followed by a flattening layer and fully-connected dense layers ('fc', green). Network output is computed through a tanh activation (purple). (**a**) This architecture was used when training only on image data. The model output is given by $\gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_{\mathrm{img}})$. (**b**) An additional dense layer was introduced when clinical features were used in addition to image data. The network output in this case is given by $\gamma_{\boldsymbol{\beta}}(\boldsymbol{x})$.

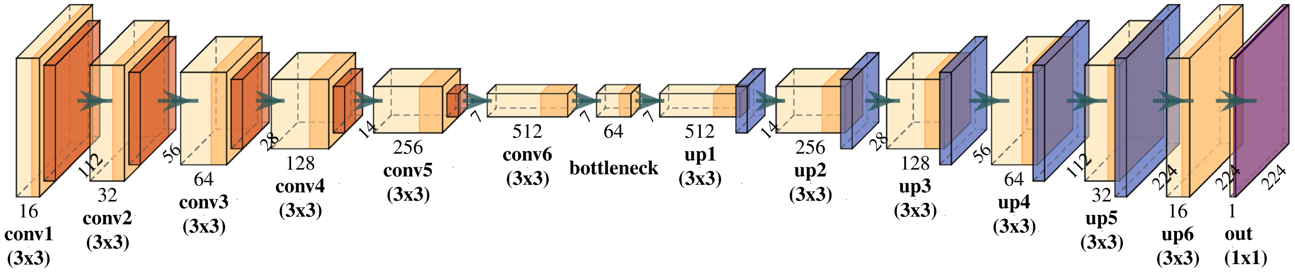

Figure 3. Architecture of the applied autoencoder. Numbers describe the shapes of computed feature maps. Convolutional layers ('conv') are comprised of convolutional filters (light orange) and Leaky ReLU ($\alpha = 0.01$) activation functions (orange). Spatial downsampling is performed using max-pooling layers (red), resulting in a set of bottleneck features. Upsampling operations ('up', blue), and convolutional layers are then used to reconstruct the input image. A sigmoid activation (purple) is used as model output to match the range of the input data.

We ran two experiments per architecture, using the last convolutional layer (denoted "last") and an earlier layer[51] (denoted by the name of the layer in the Keras implementation) of the pre-trained networks as foundation for our models. Since those models were trained on RGB images with three input channels, we renormalised each CT slice to the range [0, 255], replicated it to three channels and applied the network preprocessing functions provided by the Keras framework. All layers were fine-tuned simultaneously with a learning rate of 10^{-6} for a maximum of 20 epochs while doing early stopping (patience=5) on the internal test fold as well as reducing learning rates on plateaus (factor=0.5, patience=3, min _lr $= 10^{-7}$) via the provided Keras callbacks. A batch size of 32 was used and neither data augmentation nor weight regularisation were applied.

Deep features. Following Wang et al.[34], we trained a 2D-CNN autoencoder model that learns to reproduce input CT slices as close as possible while passing through a so called bottleneck layer which acts as a means of compression and dimensionality reduction. Successful reconstruction requires capturing of important image characteristics at the bottleneck and we assumed that relevant tumour information was also encoded within those features. The model architecture is provided in Fig. 3 and consisted of an encoder part of six convolutional layers with filter size 3×3, starting with 16 filters and doubling on each subsequent layer. Leaky ReLU ($\alpha = 0.01$) was used as activation. No BN was applied. Between convolutional layers, max-pooling was used to reduce spatial resolution by a factor of two. Finally, a last 3×3 convolutional layer with 64 filters and the same specification as above was applied to reduce the number of features in the bottleneck representation. The following decoder model was constructed as a mirror image of the encoder using upsampling layers for doubling spatial resolution in each step. The decoder's last layer was a single 1×1 convolutional filter with sigmoid activation function to produce outputs with a data range of (0, 1), matching the input image range. Using the binary-crossentropy loss function, we trained the autoencoder for 100 epochs with batches of size 32 using the AMSGrad version of the Adam optimiser with learning rate 10^{-3}. We used data augmentation by randomly

		C-index			Log-rank p value
		Exploratory cohort		Independent validation cohort	
Final activation	Batch normalisation	Training	Internal test		
3D-CNN					
tanh	Yes	0.02 (0.01–0.03)	0.39 (0.33–0.46)	**0.31** (0.22–0.39)	**0.001**
2D-CNN					
linear	No	0.02 (0.01–0.02)	0.43 (0.36–0.49)	0.39 (0.29–0.49)	0.039
linear	Yes	0.01 (0.00–0.02)	0.43 (0.36–0.49)	0.38 (0.27–0.48)	0.015
tanh	No	0.07 (0.05–0.09)	0.42 (0.36–0.48)	0.38 (0.28–0.48)	0.051
tanh	Yes	0.01 (0.01–0.02)	0.42 (0.36–0.48)	0.38 (0.27–0.48)	0.015
2D-CNN + volume					
tanh	Yes	0.06 (0.04–0.08)	0.47 (0.41–0.53)	0.40 (0.29–0.50)	0.070

Table 2. Ensemble training from scratch: C-indices for the endpoint loco-regional control (LRC) are computed by averaging the model predictions of the repeated cross-validation models to build an ensemble model. Values in parenthesis denote 95% confidence intervals. In addition, differences in LRC between Kaplan–Meier curves of the stratified patient groups are assessed by the log-rank test. Best performance is marked in bold. *C-index* concordance index, tanh, hyperbolic tangent.

shearing (shear_range = 0.1), zooming (zoom_range = 0.1) and rotating (rotation_range = 45) the input data. We then extracted the bottleneck feature maps of each slice which were of shape 7 × 7 × 64, leading to a reduction to 6.25% of the original image size (224 × 224). Those features were then flattened into a 3136 dimensional vector and a principal component analysis (PCA) was performed using the features of all slices of every patient from the training fold of the CV as a means of dimensionality reduction. Classical CPHMs were subsequently fitted on those training folds using one, two, five and ten PCA features. The learned PCA transformation was then applied to the independent validation cohort features before evaluating the performance of the trained CPHMs on those transformed features. In addition, a Lasso-based CPHM (LCPHM)[52], that automatically selects relevant features, was fit on the full set of bottleneck features of each training fold without performing a PCA for a maximum of 5000 iterations. The best hyperparameter λ, which determines the amount of L1 regularisation of the LCPHM, was obtained by another nested CV run on each training fold. This procedure was implemented using the R programming language and the glmnet package[53].

Results

Clinical model.

All available clinical features were considered to develop a clinical model for the prediction of LRC hazards. Based on the forward variable selection procedure, only the tumour volume was selected. This univariate CPHM achieved a C-index of 0.39 (95% CI: 0.32-0.45) on the exploratory cohort and a C-index of 0.39 (95% CI: 0.30–0.48) on the independent validation cohort. Stratification of the independent validation cohort into patient groups at low and high risk of loco-regional recurrence based on this clinical model showed a statistical trend approaching significance ($p = 0.052$, Supplementary Fig. 1).

Training from scratch.

An ensemble of 3D-CNNs was successfully validated for the prediction of LRC. It achieved a C-index of 0.31 (95%-CI: 0.22-0.39) on the independent validation cohort (Table 2), outperforming the clinical model. Ensembling slightly improved average single model performance (C-index: 0.32, Supplementary Table 1). Moreover, stratification of patients of the independent validation cohort (Fig. 4, top row) into groups at low and high risk of loco-regional recurrence based on the model predictions revealed significant differences in LRC ($p = 0.001$). Ensembles of 2D-CNN models trained from scratch were also successfully validated for prognosis of LRC. However, they showed higher C-indices than the 3D model (C-index: 0.38-0.39, Table 2), i.e. a performance comparable to the clinical model. Average single model performance was similar (Supplementary Table 1). All 2D ensemble models led to significant patient stratifications on the independent validation cohort for LRC or showed a statistical trend (Fig. 4, centre row) ($p = 0.051$). Table 2 also shows that the inclusion of BN and the choice of final activation did not have a strong impact on performance regarding C-indices or stratification ability of the independent validation cohort. The Spearman correlation coefficient between model predictions and z-score normalised log-tumour volume was moderate across all 30 models (with BN, tanh as final activation), with average values of 0.30 and 0.36 for the exploratory and independent validation cohort, respectively. Combining imaging data and tumour volume as network input resulted in decreased performance compared to models with only the CT image as input: a C-index of 0.40 (95%-CI: 0.29-0.50) was

Architecture	Layer name	C-index			Log-rank *p* value
		Exploratory cohort		Independent validation cohort	
		Training	Internal test		
ResNet50	last	0.06	0.37	0.39	0.17
		(0.04–0.07)	(0.31–0.42)	(0.30–0.48)	
ResNet50	activation_37	0.14	0.39	0.41	0.15
		(0.11–0.17)	(0.33–0.44)	(0.31–0.51)	
DenseNet201	last	0.05	0.39	**0.37**	0.041
		(0.04–0.06)	(0.33–0.45)	(0.27–0.47)	
DenseNet201	conv4_block48	0.12	0.43	0.43	0.032
		(0.10–0.15)	(0.37–0.50)	(0.33–0.53)	
IRNV2	last	0.08	0.38	0.41	0.25
		(0.06–0.10)	(0.32–0.44)	(0.31–0.52)	
IRNV2	block17_10_ac	0.26	0.41	0.42	**0.023**
		(0.22–0.31)	(0.36–0.47)	(0.32–0.53)	

Table 3. Ensemble of transfer learning models: C-indices for the endpoint loco-regional control (LRC) are computed by averaging the model predictions of the repeated cross-validation models to build an ensemble model. Values in parenthesis denote 95% confidence intervals. In addition, differences in LRC between Kaplan–Meier curves of the stratified patient groups are assessed by the log-rank test. Best performance is marked in bold. *C-index* concordance index, *IRNV2* inceptionResNetV2.

obtained on the independent validation cohort and model predictions did not result in a statistically significant stratification ($p = 0.070$).

Transfer learning. For transfer learning, the ensemble of DenseNet201 models in combination with its last convolutional layer as the foundation was successfully validated for prognosis of LRC and achieved the best C-index of 0.37 (95%-CI: 0.27-0.47) on the independent validation cohort (Table 3), which was slightly better than the clinical model. Compared to average single model peformance (C-index: 0.41, Supplementary Table 2) this was an improvement of 0.04. Moreover, a statistically significant stratification into low and high risk groups of loco-regional recurrence was achieved by this ensemble for the independent validation cohort (Fig. 4, bottom row) ($p = 0.041$). Using the last convolutional layer as foundation, ensembles of ResNet50 or IRNV2 models were not able to successfully stratify patients of the independent validation cohort. Layers different from the last convolutional layer of the pre-trained models as input for the newly added dense layers resulted in slightly worse C-indices in all cases.

Boxplots showing the variability of ensemble predictions for patients of the independent validation cohort are provided in Supplementary Figs. 2, 3 and 4 for the ensemble of 3D-CNN models, 2D-CNN models and DenseNet201 models, respectively.

Deep features. The prognostic performance of classical CPHMs using bottleneck features of autoencoder models as covariates are given in Table 4. Model performance was inferior to the clinical model in all scenarios and none of the models achieved a statistically significant stratification of the independent validation cohort into low and high risk groups. The best C-index on the independent validation cohort was 0.42 (95%-CI: 0.32–0.53), obtained by the LCPHM ensemble. The ensemble model improved the C-index on the independent validation cohort by 0.03 compared to the average single model C-index (Supplementary Table 3). The amount of the full variance of the data captured by the PCA features is provided in Supplementary Table 4.

Discussion

We investigated deep learning methods in a survival analysis setting for the endpoint LRC, based on treatment-planning CT images of locally advanced HNSCC patients treated with primary radiochemotherapy. Best performance and successful validation was achieved by an ensemble of 3D-CNNs with a C-index of 0.31 on the independent validation cohort. Patient risk groups defined by the model predictions showed significant differences in LRC ($p = 0.001$). Ensembles of different 2D-CNN approaches performed similar to a clinical CPHM based on the tumour volume (independent validation C-index of 0.39). Compared to using only a single trained model instance, our analysis revealed benefits in using model ensembles for final predictions, which is in line with the reasoning of Dietterich[43].

Overall, reported performances for 2D-CNNs were comparable to results previously published from our group by Leger et al.[17]. They evaluated multiple combinations of feature selection algorithms and classical machine learning models based on handcrafted radiomics features on the same dataset. An average independent validation C-index over all combinations of 0.62 was achieved (which corresponds to a C-index of 0.38 in our context, as explained in the "Methods" section). Similarly, Haarburger et al.[28] reported C-indices between 0.585

	C-index			**Log-rank *p* value**
	Exploratory cohort			
Feature selection + ML algorithm	**Training**	**Internal test**	**Independent validation cohort**	
- + LCPHM	0.01	0.50	**0.42**	**0.19**
	(0.00–0.01)	(0.43–0.57)	(0.32–0.53)	
PCA(1) + CPHM	0.49	0.53	0.54	0.63
	(0.42–0.56)	(0.47–0.60)	(0.42–0.66)	
PCA(2) + CPHM	0.47	0.51	0.53	**0.19**
	(0.40–0.54)	(0.44–0.58)	(0.42–0.64)	
PCA(5) + CPHM	0.44	0.50	0.50	0.72
	(0.37–0.50)	(0.44–0.57)	(0.41–0.60)	
PCA(10) + CPHM	0.35	0.42	0.43	0.40
	(0.29–0.40)	(0.36–0.48)	(0.33–0.53)	

Table 4. Ensemble of autoencoder models: C-indices for the endpoint loco-regional control (LRC) are computed by averaging the model predictions of the repeated cross-validation models to build an ensemble model. Values in parenthesis denote 95% confidence intervals. In addition, differences in LRC between Kaplan–Meier curves of the stratified patient groups are assessed by the log-rank test. Best performance is marked in bold. *C-index* concordance index, *ML* machine learning, *CPHM* Cox proportional hazards model, *LCPHM* Lasso-Cox proportional hazards model, *PCA* principal component analysis.

and 0.623 using a CNN based CPHM on a CT imaging dataset of lung cancer patients. However, they suggested to reformulate the regression problem as a classification task. This was due to their GPU memory limitations which did not allow large enough batch sizes for good approximations of the partial log-likelihood function of the CPHM. We did not observe problems with small batch sizes (see Supplementary Table 5), but investigated their approach in an additional analysis: We evaluated ensemble model performance of 2D-CNNs using a cutoff of 24 months for LRC. Samples of 72 patients previously used in our analysis had to be discarded due to censoring before the cutoff value, clearly demonstrating the downside of binarising the time-to-event variable. We achieved an area under the receiver operating characteristic (AUROC) curve of 0.60 on the independent validation cohort which is comparable to the reported AUROC values of 0.598 and 0.636[28].

We also provided empirical evidence (Supplementary Table 5) that approximation of the Cox partial log likelihood by batches does not seem to be problematic since models were able to learn well on the training set regardless of small (32) or large (256) batch sizes. However, we did not observe additional benefits for increased batch sizes. Moreover, we found that adding more samples of a single patient for training and inference did not further improve 2D-CNN results (Supplementary Table 5).

Using deep autoencoder features turned out to be the least effective approach among our investigations, since it never achieved statistically significant patient stratifications and showed the worst C-indices on the independent validation cohort. This might be due to having used a too large compression in our network design of the bottleneck features or having chosen a too small amount of PCA features. This is indicated by the performance improvements when switching from five to 10 PCA features, as well as by results reported by Wang et al.[34]. There, using 16 autoencoder features (of a different network architecture and without PCA), C-indices of 0.713 and 0.694 on two cohorts diagnosed with high-grade serous ovarian cancer were reported.

Most CNN models struggled with overfitting, as can be seen in the large discrepancies of C-indices between training and internal test/independent validation (Tables 2, 3), which is also reflected in the large separation of KM curves between low and high risk groups in the training column of Fig. 4. Adding multiple regularisation approaches such as L1 and L2 weight regularisation, increased dropout rates and data augmentation to the fitting procedure of our 2D-CNN models trained from scratch, we observed indeed drops in training performance for most approaches but without improvements on the independent validation, Supplementary Table 6. Similar observations were made for the 3D-CNN models. There, we employed a 3D data augmentation strategy using code from https://github.com/MIC-DKFZ/batchgenerators which included elastic deformations (deformation scale (0, 0.25)), random rotations in the range of [– 15, 15] degrees for each of the three spatial axes, random rescaling in the range of (0.75, 1.25), mirroring, random brightness multiplications in the range (0.7, 1.5), gaussian noise additions with noise variance parameter set to (0, 0.05) and per channel gamma transformations using a gamma range parameter of (0.5, 2). However, ensemble results were similar (independent validation C-index 0.30, log-rank $p = 0.003$) to the results obtained without using data augmentation. Throughout our experiments, we also observed C-index performance on the internal test folds of the CV to have much higher variance compared to results of the training folds and the independent validation. We attribute this to the small sample sizes of about 20 patients in the internal test folds during the 10-fold CVs. Ensemble predictions for the internal test patients were also consistently worse than the ensemble predictions for the patients of the independent validation cohort, which might be due to the smaller number of models used for building the ensembles (only three models for the former but 30 for the latter) and inherent statistical differences between patients of the exploratory and the independent validation cohort for, e.g. tumour volume or tumour site.

The performance benefits observed for 3D-CNNs may have multiple causes. Firstly, those models allowed to incorporate potentially relevant spatial (three dimensional) context around the tumour during training. In

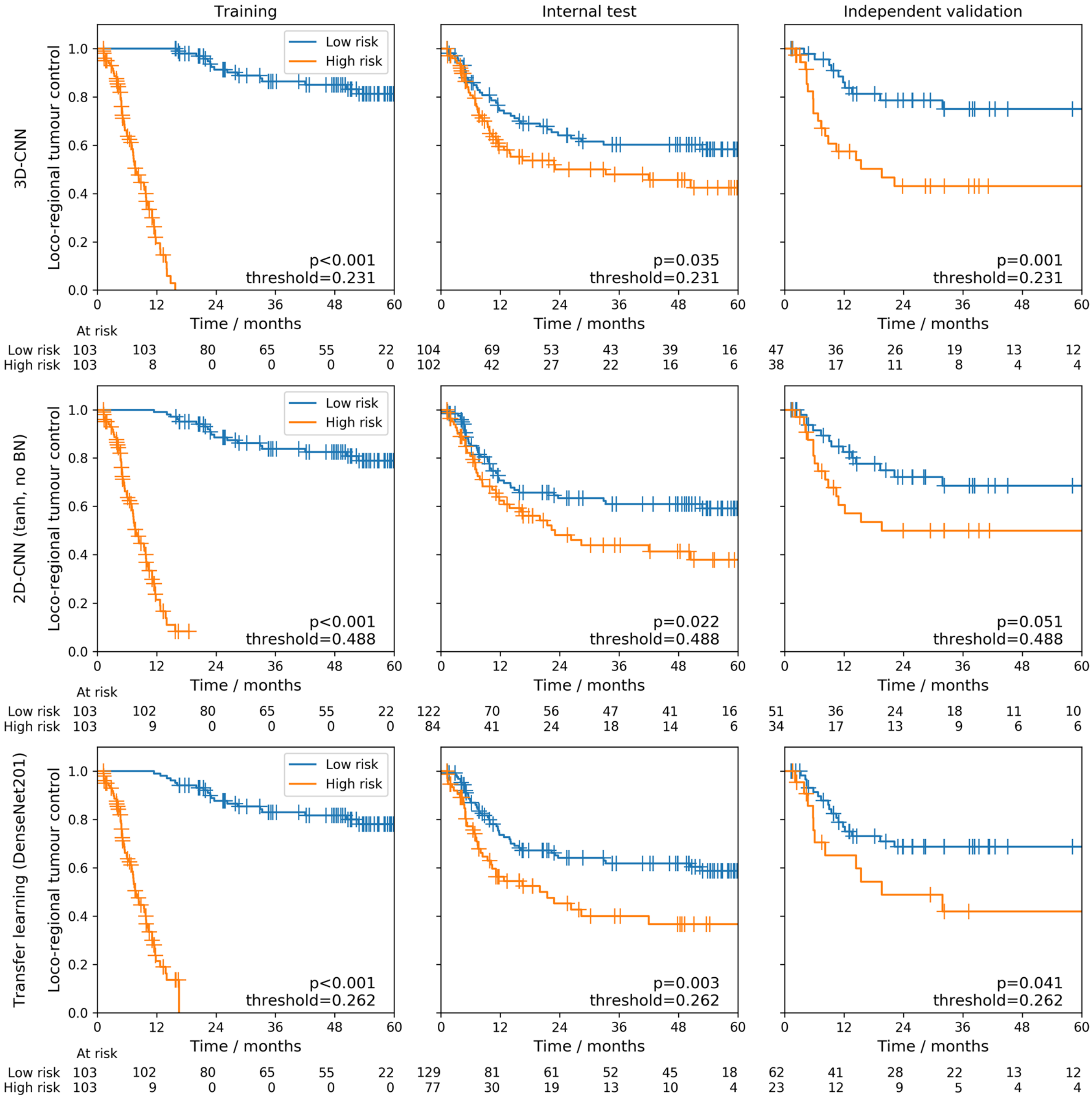

Figure 4. Ensemble Kaplan–Meier curves: Kaplan–Meier curves for patient groups at low risk (blue) and high risk (orange) of loco-regional recurrence for training and internal test folds as well as for the independent validation cohort. The stratification was created using the median of the training ensemble predictions as cutoff. The top row shows the curves obtained from an ensemble of 3D-CNN models trained from scratch based on the architecture of Hosny et al.[23] with tanh as final activation. The centre row shows the curves obtained from an ensemble of 2D-CNN models trained from scratch without batch normalisation and tanh as final activation. The bottom row shows the curves obtained from an ensemble of transfer learning models based on DenseNet201 with the last convolutional layer as foundation.

contrast, 2D-CNN models by design were not capable of exploiting this additional information during training, but only during inference, by considering predictions of multiple patient slices. Secondly, the input image size differed between 2D and 3D-CNN models. While 2D-CNNs analysed the full axial plane, 3D-CNN models processed only a relatively small axial area close to the tumour which might have allowed them to learn more relevant tumour features and not to get distracted by possibly uninformative image regions. Thirdly, the model's architecture and hyperparameters differed between 2D and 3D-CNN models which could have influenced the observed performance differences.

Our analysis contains some limitations: Our data set, even though competitive in size in the field of medical imaging, might be too small to obtain better results. For 2D-CNN models, even transfer learning was not able to circumvent this limitation, which might also be due to the large translational gap that exists between natural RGB images and CT scans. Federated learning[54,55] seems to be a promising way to tackle the small sample size problem of medical imaging. This includes setting up infrastructures to allow to collaboratively train models on data of multiple institutions without violating data-privacy regulations. Also, exploring generative adversarial networks for enhancing dataset sizes through simultaneous generation of synthetic image samples and plausible time-to-event labels[56,57] might provide a potentially interesting task. However, for HNSCC, treatment-planning

CT scans may simply not contain much more predictive information to achieve better performance, no matter the deep learning approach, model architecture or hyperparameters. As previously indicated[3,9,58], considering additional imaging during the course of treatment or additional imaging modalities such as MRI or PET may offer improved predictive potential. Another limitation of our analysis concerns the Cox partial log-likelihood function, as given by equation (1), which does not account for ties in the data. This can very well occur if multiple samples of the same patient are present in a single training batch. Therefore, we plan on using e.g. Efrons correction method[59] in future analysis but refrained from that in our current experiments in order to avoid introduction of additional complexity in the loss function. Instead, we experimented with using slight random perturbations on the observed event times to avoid exact matches. We did, however, not observe noteworthy changes in model performance (see first row of Supplementary Table 6). An alternative to the CPHM is the combination of deep learning with accelerated failure time models, as demonstrated by Chapfuwa et al.[57] on clinical data. Due to their fully-parametric nature, direct prediction of event times becomes easier and non-monotonic hazard functions can be modelled.

Deep learning approaches on treatment-planning CT images can be useful building blocks on the way to achieve the goal of personalisation of radiotherapy. They may be extended using additional information, e.g. from tumour histology or molecular samples. Nevertheless, deep learning approaches should not be considered the universal remedy since they also bring with them some drawbacks compared to simpler models. Those include increased computational complexity and difficulties in understanding the image-based causes of their predictions, leading to decreased model interpretability.

In this study, we implemented CNNs for the prediction of LRC after primary radiochemotherapy of locally advanced HNSCC based on CT imaging. An ensemble of 3D-CNN models was successfully validated and showed an improved performance compared to 2D-CNN approaches and a clinical model. Risk groups defined on these models differed significantly in LRC. In the future, we aim to assess robustness and translational ability of our trained models by applying them to data of the prospective HNPrädBio trial of the DKTK-ROG as another independent validation (NCT02059668)[60].

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request. Experimental output and trained models are accessible from http://doi.org/10.14278/rodare.255. Python and R code of our analyses is available from https://github.com/oncoray/cnn-hsncc.

Received: 27 February 2020; Accepted: 20 July 2020
Published online: 24 September 2020

References

1. Baumann, M. *et al.* Radiation oncology in the era of precision medicine. *Nat. Rev. Cancer* **16**, 234–249 (2016).
2. Leemans, C. R., Braakhuis, B. J. M. & Brakenhoff, R. H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **11**, 9–22 (2011).
3. Zips, D. *et al.* Exploratory prospective trial of hypoxia-specific PET imaging during radiochemotherapy in patients with locally advanced head-and-neck cancer. *Radiother. Oncol.* **105**, 21–28 (2012).
4. Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 1–9 (2014).
5. Lohaus, F. *et al.* HPV16 DNA status is a strong prognosticator of loco-regional control after postoperative radiochemotherapy of locally advanced oropharyngeal carcinoma: Results from a multicentre explorative study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). *Radiother. Oncol.* **113**, 317–323 (2014).
6. Vallières, M., Freeman, C. R., Skamene, S. R. & Naqa, I. E. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60**, 5471–5496 (2015).
7. Linge, A. *et al.* HPV status, cancer stem cell marker expression, hypoxia gene signatures and tumour volume identify good prognosis subgroups in patients with HNSCC after primary radiochemotherapy: A multicentre retrospective study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). *Radiother. Oncol.* **121**, 364–373 (2016).
8. Linge, A. *et al.* Low cancer stem cell marker expression and low hypoxia identify good prognosis subgroups in HPV(-) HNSCC after postoperative radiochemotherapy: A multicenter study of the DKTK-ROG. *Clin. Cancer Res.* **22**, 2639–2649 (2016).
9. Löck, S. *et al.* Residual tumour hypoxia in head-and-neck cancer patients undergoing primary radiochemotherapy, final results of a prospective trial on repeat FMISO-PET imaging. *Radiother. Oncol.* **124**, 533–540 (2017).
10. Bogowicz, M. *et al.* Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. *Radiother. Oncol.* **125**, 385–391 (2017).
11. Vallières, M. *et al.* Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci. Rep.* **7**, 10117 (2017).
12. Schmidt, S. *et al.* Development and validation of a gene signature for patients with head and neck carcinomas treated by postoperative radio(chemo)therapy. *Clin. Cancer Res.* **24**, 1364–1374 (2018).
13. Deist, T. M. *et al.* Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med. Phys.* **45**, 3449–3459 (2018).
14. Parmar, C. *et al.* Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front. Oncol.* **5**, 272 (2015).
15. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
16. Li, Q. *et al.* A fully-automatic multiparametric radiomics model: towards reproducible and prognostic imaging signature for prediction of overall survival in glioblastoma multiforme. *Sci. Rep.* **7**, 1–9 (2017).
17. Leger, S. *et al.* A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci. Rep.* **7**, 13206 (2017).
18. Peeken, J. C. *et al.* Radiomics in radiooncology: challenging the medical physicist. *Phys. Med.* **48**, 27–36 (2018).
19. Sollini, M., Antunovic, L., Chiti, A. & Kirienko, M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur. J. Nucl. Med. Mol. Imaging* **46**, 2656–2672 (2019).
20. Ger, R. 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* **14**, e0222509 (2019).

21. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 234–241 (Springer Verlag, 2015).
22. Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G. & Seuntjens, J. Deep learning in head & neck cancer outcome prediction. *Sci. Rep.* **9**, 1–10 (2019).
23. Hosny, A. *et al.* Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS Med.* **15**, 1–25 (2018).
24. Baek, S. *et al.* Deep segmentation networks predict survival of non-small cell lung cancer. *Sci. Rep.* **9**, 17286 (2019).
25. Meyer, P., Noblet, V., Mazzara, C. & Lallement, A. Survey on deep learning for radiotherapy. *Comput. Biol. Med.* **98**, 126–146 (2018).
26. Zhu, X., Yao, J. & Huang, J. Deep convolutional neural network for survival analysis with pathological images. In *Proceedings—2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016* 544–547 (2017).
27. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. In *Proceedings of the National Academy of Sciences of the United States of America* (2018).
28. Haarburger, C., Weitz, P., Rippel, O. & Merhof, D. Image-based survival analysis for lung cancer patients using CNNs (2018). arXiv:1808.09679v1.
29. Katzman, J. L. *et al.* DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network (2018). arXiv:1606.00931v3.
30. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **14**, 1–18 (2018).
31. Liu, R. *et al.* Exploring deep features from brain tumor magnetic resonance images via transfer learning. *Proc. Int. Jt. Conf. Neural Netw.* **2016–Octob**, 235–242 (2016).
32. Paul, R. *et al.* Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography* **2**, 388 (2016).
33. Lao, J. *et al.* A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports* **7**, 1–8 (2017).
34. Wang, S. *et al.* Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother. Oncol.* **132**, 171–177 (2019).
35. Linge, A. *et al.* Independent validation of tumour volume, cancer stem cell markers and hypoxia-associated gene expressions for HNSCC after primary radiochemotherapy. *Clin. Transl. Radiat. Oncol.* **16**, 40–47 (2019).
36. Welz, S. *et al.* Prognostic value of dynamic hypoxia PET in head and neck cancer: Results from a planned interim analysis of a randomized phase II hypoxia-image guided dose escalation trial. *Radiother. Oncol.* **124**, 526–532 (2017).
37. Harrell, F. E., Lee, K. L. & Mark, D. B. Tutorial in biostatistics multivariable prognostic models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
38. Mayr, A. & Schmid, M. Boosting the concordance index for survival data: a unified framework to derive and evaluate biomarker combinations. *PLoS ONE* **9**, e84483 (2014).
39. Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P. & Raykar, V. C. On ranking in survival analysis: bounds on the concordance index. *Adv. Neural Inf. Process. Syst.* **20**, 1209–1216 (2008).
40. Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?. *Bioinformatics* **24**, 2200–2208 (2008).
41. Schröder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**, 3206–3208 (2011).
42. Pencina, M. J. & DAgostino, R. B. Overall C as a measure of discrimination in survival analysis model specific population value and confidence interval estimation. *Stat. Med.* **23**, 2109–2123 (2004).
43. Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 1–15 (Springer, Berlin, 2000).
44. Zwanenburg, A. *et al.* Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* **9**, 614 (2019).
45. Chollet, F. *et al.* Keras. https://keras.io (2015).
46. Reddi, S.J., Kale, S. & Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations* (2018).
47. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556.
48. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition (2015). arXiv:1512.03385.
49. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks (2016). arXiv:1608.06993.
50. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning (2016). arXiv:1602.07261.
51. Mormont, R., Geurts, P. & Maree, R. Comparison of deep transfer learning strategies for digital pathology. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2018-June, 2343–2352 (IEEE Computer Society, 2018).
52. Tibshirani, R. The LASSO method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
53. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Coxs proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
54. Jochems, A. *et al.* Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. *Radiother. Oncol.* **121**, 459–467 (2016).
55. Jochems, A. *et al.* Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. *Int. J. Radiat. Oncol. Biol. Phys.* **99**, 344–352 (2017).
56. Aggarwal, K., Kirchmeyer, M., Yadav, P., Keerthi, S.S. & Gallinari, P. Regression with conditional GAN (2019). arXiv:1905.12868.
57. Chapfuwa, P. *et al.* Adversarial time-to-event modeling. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, 735–744 (2018).
58. Leger, S. *et al.* CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer. *Radiother. Oncol.* **130**, 10–17 (2019).
59. Efron, B. The efficiency of Coxs likelihood function for censored data. *J. Am. Stat. Assoc.* **72**, 557–565 (1977).
60. Observational Study on Biomarkers in Head and Neck Cancer (HNprädBio). ClinicalTrials.gov [Internet], Bethesda (MD): National Library of Medicine (US), 2000 Feb 29. Identifier NCT02059668 [registered 2014 Feb 11, updated 2019 Feb 18, cited 2020 Jan 30]. https://clinicaltrials.gov/ct2/show/NCT02059668 (2014).

Acknowledgements

We acknowledge the creators of the github repository at https://github.com/HarisIqbal88/PlotNeuralNet/ for providing code for the creation of neural network architecture plots. We used their code for creating Figs. 2 and 3. Also we would like to thank the authors of the repository at https://github.com/MIC-DKFZ/batchgenerators for making their data augmentation code publically available. “Neural network” icon used in Fig. 1 by Ben Davis from thenounproject.com (https://thenounproject.com/term/neural-network/1718441/). "Machine learning" icon used in Fig. 1 by Mohamed Mb from thenounproject.com (https://thenounproject.com/search/?q=machine%20learning&i=1705417). The author SLe is supported by the Federal Ministry of Education and Research (BMBF-13GW0211D).

Author contributions

S.S. developed analysis tools and, together with S.L., A.Z. and S.L. analysed the data and wrote the paper. M.B., M.K. and S.L. conceived of the Project and reviewed the manuscript. K.L., E.G.C.T. and F.L. provided expert guidance, data and reviewed the manuscript. A.L., A.S., G.K., I.T., N.G., M.G., P.B., J.v.d.G., U.G., C.B., J.C.P., S.E.C., S.B., D.Z. and C.R. provided data and reviewed the manuscript.

Funding

Open access funding provided by Projekt DEAL.

Competing interests

Dr. Linge is involved in an ongoing publicly funded (German Federal Ministry of Education and Research) Project with the companies Medipan, Attomol GmbH, GA Generic Assays GmbH, Gesellschaft für medizinische und wissenschaftliche genetische Analysen, Lipotype GmbH and PolyAn GmbH (2019–2021). For the present manuscript, Dr. Linge confirms that this above mentioned funding source was not involved. Dr. Richter received individual funding as lecturer from Siemens Healthineers (2018). OncoRay has institutional research agreements with Siemens Healthineers. Furthermore, OncoRay has an institutional agreement as reference center for dual-energy CT in radiotherapy as well as a software evaluation contract with Siemens Healthineers. In the past 5 years, Dr. Krause received funding for her research Projects by IBA (2016), Merck KGaA (2014–2018 for preclinical study; 2018-2020 for clinical study), Medipan GmbH (2014–2018). In the past 5 years, Dr. Troost, Dr. Krause and Dr. Löck have been involved in an ongoing publicly funded (German Federal Ministry of Education and Research) Project with the companies Medipan, Attomol GmbH, GA Generic Assays GmbH, Gesellschaft für medizinische und wissenschaftliche genetische Analysen, Lipotype GmbH and PolyAn GmbH (2019–2021). For the present manuscript, none of the above mentioned funding sources were involved. In the past 5 years, Dr. Baumann attended an advisory board meeting of MERCK KGaA (Darmstadt), for which the University of Dresden received a travel grant. He further received funding for his research Projects and for educational grants to the University of Dresden by Teutopharma GmbH (2011–2015), IBA (2016), Bayer AG (2016–2018), Merck KGaA (2014–2030), Medipan GmbH (2014-2018). For the German Cancer Research Center (DKFZ, Heidelberg) Dr. Baumann is on the supervisory boards of HI-STEM gGmbH (Heidelberg). Dr. Baumann, as former chair of OncoRay (Dresden) and present CEO and Scientific Chair of the German Cancer Research Center (DKFZ, Heidelberg), was or is responsible for collaborations with a multitude of companies and institutions, worldwide. In this capacity, he has signed/signs contracts for his institute(s) and for the staff for research funding and/or collaborations with industry and academia, worldwide. In this role, he was/is further responsible for commercial technology transfer activities of his institute(s), including patent and other similar IP portfolios. Dr. Baumann confirms that none of the above funding sources were involved in the preparation of this paper. University Department of Radiation Oncology Tübingen receives financial and technical support from Elekta AB (Stockholm, Sweden) under a research agreement. The other authors have nothing to disclose.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70542-9.

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.