university of groningen / faculty of science and engineering / biomedical engineering

Improving the Prediction of Radiation-Induced Taste Loss in Head and Neck Cancer

Hendrike Neh

S4267621

Department of radiation oncology, University Medical Centre Groningen

Period: 07/02/2022 - 05/08/2022

Master's project

1st Examiner: dr. ir. P.M.A. (Peter) van Ooijen, Radiation Oncology department, University Medical Center Groningen

2nd Examiner: L. V. (Sanne) van Dijk, PhD, Radiation Oncology department, University Medical Center Groningen

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
CITOR	Comprehensive Individual Toxicity Risk
CNN	Convolutional Neural Network
Crico	Cricopharyngeal Muscle
CT	Computed Tomography
DCNN	Deep Convolutional Neural Network
DLC	Deep Learning Contouring
DVH	Dose-Volume Histograms
FN	False Negative
FoR UID	Frame of Reference Unique Identifier
FP	False Positive
FU	Follow Up
HL	Hosmer-Lemeshow Test
HNC	Head and Neck Cancer
HPV	Human Papilloma Virus
IMPT	Intensity Modulated Proton Therapy
IMRT	Intensity Modulated Radiation Therapy
MRI	Magnetic Resonance Imaging
NTCP	Normal Tissue Complication Probability
OAR	Organs At Risk
OR	Odds Ratio
PCM	Pharyngeal Constrictor Muscle
PET-CT	Positron Emission Tomography - CT
rCNN	Residual Convolutional Neural Network
SCC	Squamous Cell Carcinoma
SD	Standard Deviation
SGD	Stochastic Gradient Descent
TN	True Negative
TP	True Positive
UMCG	University Medical Center Groningen
VMAT	Volumetric Modulated Arc Therapy

Abstract

Purpose:

Taste loss is a common side effect of head and neck cancer (HNC) radiotherapy treatment, and its prediction is important to increase the health and quality of life of survivors. This project aimed to improve the prediction of late taste loss at six months after treatment compared to previously developed logistic regression normal tissue complication probability (NTCP) models for patients undergoing radiation therapy for HNC by including a tongue mucosa structure in existing conventional NTCP models as well as developing a deep learning based NTCP model.

Materials and Methods:

Included patients with HNC were treated with radiotherapy between 2007 and 2022. The baseline model was derived from the NTCP model by Van den Bosch et al. with oral cavity mean dose, parotid gland mean dose and age as parameters of the logistic regression model. The new tongue mucosa structure was derived from the existing oral cavity structure and their mean dose performances in the logistic regression model were compared by performing a univariate analysis and model refit. A deep convolutional neural network (DCNN) and residual convolutional neural network (rCNN) were trained systematically using CT, organ segmentation and 3D dose distribution as input.

Results:

Of 949 included patients with HNC and available endpoint data, 26.5% (n = 252) patients reported moderate-severe taste loss at 6 months post treatment. A univariable analysis showed that the oral cavity mean dose was a more important predictor than the tongue mucosa mean dose, but that the tongue mucosa was still preferred over the parotid gland mean dose and age at treatment. The logistic regression model with the tongue mucosa mean dose (AUC: 0.717; calibration slope: 1.02) did not perform better than the reference model with oral cavity mean dose (AUC: 0.724; calibration slope: 1.01). The DCNN (AUC: 0.682) and rCNN (AUC: 0.684) both performed worse than the reference model (AUC: 0.692) on the test set.

Conclusion:

This study presented two different NTCP models based on a new tongue mucosa structure and deep learning-based approach to predict taste loss in HNC patients at 6 months post treatment. No improvement to existing models was achieved and more work must be done to optimize the techniques used.

1. Introduction

Head and neck cancer (HNC) is among the ten most common types of cancer worldwide (*1, 2*). HNC includes different cancer locations in the upper aerodigestive tract such as oral cavity, oropharynx and larynx (*3*). In 2019 more than 3100 patients were diagnosed with HNC in the Netherlands (*4*). Radiotherapy, either in combination with surgery or chemotherapy or stand-alone, is a pivotal treatment of HNC; with around 2000 patients treated with radiotherapy in the Netherlands yearly. In recent years, the number of relatively young HNC survivors is growing. More intensified treatment regimens like fractionation and intensity-modulated radiotherapy (IMRT) have been developed that resulted in improved overall survival (*5–7*). Furthermore, the incidence of human papilloma virus (HPV)-related tumours has increased, especially for HPV associated oropharyngeal cancer which have more favourable outcomes (*8–10*). Survivors may suffer from side effects years after treatment that gravely impact their quality of life (*11, 12*). This stresses the importance of preventing late radiation-induced toxicities, which may persist or occur years after treatment.

Patients undergoing radiotherapy treatment for head and neck cancer are likely to experience taste loss or taste alteration as a side effect (*13, 14*). Being able to taste has three main purposes: pleasure, defence, and sustenance. When deprived of this sense, malnutrition and gastrotomy tube dependency, as well as a general lower quality of life can be the consequence (*13*). Taste alteration has a huge impact on the patients quality of life and is among the 5 highest patient scored symptom scores (*15*) and shows a maximum around 6 months after treatment (*15, 16*). Predicting taste loss prior to treatment in HNC patients is important to increase the health and quality of life of survivors.

Side effects and the general risk of radiation-induced toxicity are commonly predicted by normal tissue complication probability (NTCP) models. Their prediction can be used to choose treatment techniques or adjust treatment plans to achieve a more favourable outcome. Conventional NTCP models are for example used in current practice when comparing the radiation plans of photon and proton irradiation techniques. An NTCP model for all major HNC radiation therapy related toxicities was developed in recent years by Van den Bosch et al. (*17*). Their model for late taste loss at six months had an area under the curve (AUC) of 0.68 on their development cohort with parotid gland mean dose, oral cavity mean dose and age as predictors. However, the restriction to mean dose values of conventionally used organ at risk (OAR) structures does not allow the NTCP to sufficiently model the large variation between patients.

A possible solution for more accurate prediction was introduced by Stieb et al. (*18*) suggesting that the dose given to the taste bud bearing tongue mucosa might be an alternative predictor to the oral cavity for late taste loss in HNC patients. Furthermore, in line with the current improvements and research in this field, the use of a deep learning model to achieve a toxicity prediction based on 3D voxel-wise distributions of dose as well as three dimensional CT images might allow for an NTCP model based on more detailed and less global predictions.

This project aimed to improve the prediction of late taste loss at six months after treatment compared to previously developed logistic regression NTCP models for patients undergoing radiation therapy for HNC by including a tongue mucosa structure in existing conventional NTCP models as well as developing a deep learning based NTCP model.

2. Background

2.1 Radiation therapy

Radiation Therapy, next to chemotherapy or surgery, is one of the common treatment options for patients with cancer. It is based on the damaging effect that ionizing radiation has on tissue cells. The radiation causes DNA breaks to be induced that prohibit a cell from dividing successfully if not repaired in time. A cell that cannot divide successfully due to damaged DNA will die. Since cancerous cells divide very fast, they have a higher radiosensitivity than normal tissue and this effect can be utilized in radiotherapy.

The prescribed radiation dose is given in smaller fractions over multiple days allowing the normal tissue to repair damaged parts of DNA while the faster dividing cancer cells die. Still, the dose given to normal tissue and especially radiosensitive OAR must be minimized as much as possible to avoid toxicities. Recent improvements in photon beam treatment techniques include volumetric modulated arc therapy (VMAT). Compared to conventional techniques with a single or a few beam angles, this newer method delivers the radiation dose continuously as the treatment machine rotates. This allows for closer target coverage while simultaneously improving the sparing of OARs (*19*).

A treatment plan is created for each patient prior to radiotherapy. MRI and PET-CT images are used as guidance when the tumour volume and OARs are delineated on the planning CT. In a designated software, constraints can be given to all available structures. These constraints will for example often be a minimum dose to the target volume and a maximum dose to OAR. A computer algorithm, often a Monte Carlo Simulation, will determine the most optimal treatment plan based on the given constraints. The treatment plan can be accepted by the treating clinicians once all constraints are met.

2.2 Deep learning models in medicine

Deep learning is a fast-growing field of innovation and research that is being applied to all areas of our everyday life. It allows computational models with multiple processing layers to learn representations of data using the method of abstraction (*20*). From basic image classification tasks to autonomous driving, deep learning methods are able to deliver state-of-the-art performance. The medical field has picked up on the large potential in recent years and many research groups have focussed their interest on this topic. The publication by Egger et al. (*21*) summarizes the findings from all PubMed review articles published between 2017 and 2019 on deep learning in the medical field. It demonstrates the variety in applications and increased interest across all medical domains. The potential in the field of radiation oncology is specifically demonstrated by Boldrini et al. (*22*). They conclude that deep learning, when implemented in a hospital setting, may aid clinicians when predicting treatment outcomes and toxicities, as well as allowing for fast and robust segmentation.

2.3 Outcome and toxicity prediction using deep learning

Appelt et al. (*23*) gives an overview of ten published papers from recent years using deep learning methods for outcome prediction including 3D dose information. The discussed studies lay the groundwork for the technical application of convolutional neural networks-based models that input imaging and clinical data to predict oncologic outcomes following radiotherapy. Additionally, these could lead to a better understanding of spatial variation in radiosensitivity. Nevertheless, these studies are challenged by the small number of patient data available and the lack of external validation. Zhen et al. (*24*) published a study on utilizing a deep convolutional neural network (DCNN) to predict rectum toxicity in cervical cancer patients treated with radiotherapy. They achieved satisfactory prediction results using a transfer learning approach. With this approach, a pre-trained model that was developed on a large patient cohort is tuned to fit a new application. In this way the common issue

with limited patient data when training a DCNN from scratch can be overcome. Another study by Ibragimov et al. (*25*) developed a convolutional neural network (CNN) that recognized the importance of sparing certain normal tissue regions to avoid radiation-induced toxicity. Only CT images and dose distributions were given as input and no organ segmentations.
The study most related to the research presented in this thesis was published by Men et al. (*26*) in 2019. The authors trained a deep learning model for predicting xerostomia due to radiation therapy in head and neck cancer patients. The data of 784 patients treated with photon radiation therapy for squamous cell carcinoma was used, out of which 279 patients were categorized as toxicity cases and 505 patients as non-toxicity cases. The planning CT, 3D dose distribution, and segmentation contours of the parotid and submandibular glands were used as input for a residual CNN. Better results than with their own logistic regression model were achieved.

2.4 NTCP model

To predict or estimate the risk of radiation-induced complication and toxicity, normal tissue complication probability (NTCP) models are often used. An NTCP model calculates the probability that a given dose of radiation will cause an organ or structure to experience complications. According to Van den Bosch et al. (*17, 27*) suitable NTCP models, that contain the most relevant OAR with reliable dose-response estimates, were lacking, leading to their own study on developing a comprehensive toxicity risk profiling for head and neck cancer. Because their NTCP model will be used as a comparison model for this research, I will briefly describe their methods and findings.
The comprehensive individual toxicity risk (CITOR) profile consists of NTCP models for 22 common toxicities related to head and neck cancer treatment at 10 time points during and after treatment. Multiple individual models that predict the toxicity risk for each toxicity and each time point individually were developed and then combined to create a single output. Their goal was to accurately model dose-response relationships that can be used to individualize treatment optimization resulting in the lowest overall toxicity burden. In order to create the model initial candidate predictors were picked based on prior knowledge and clinical expertise. This was done per toxicity domain. Taste loss falls under the domain of salivary toxicity and the initial predictors picked were mean dose to parotid glands, submandibular glands, oral cavity and buccal mucosa, integral dose, age, neck irradiation, treatment modality, tumour site, baseline toxicity, volume of the parotid glands and volume of the submandibular glands. The predictor selection was then done by univariable analysis. The final parameters picked as inputs for the logistic regression model for taste toxicity at 6 months can be seen in Table 1.
Logistic regression models are commonly used to predict a binary outcome given a set of input parameters. This is done according to Formula (1), where β_0 is the intercept and β_i is the regression coefficient multiplied by a predictor value x_i. In the case of taste loss prediction in the Van den Bosch et al. study, the three parameters are age, parotid mean dose and oral cavity mean dose. Their regression coefficient values are shown in Table 1. More information on the materials and methods used can be found in the publications by Van den Bosch et al. (*17, 27*).

$$p = \frac{1}{1 + e^{(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}} \quad (1)$$

Table 1: Regression coefficients used for the logistic regression model by Van den Bosch et al. with parotid mean dose, oral cavity mean dose and age as predictors. (a: ln(mean dose to both parotid glands); b: sqrt(mean dose to oral cavity))

	Regression coefficient
Parotid mean dose (β_1)[a]	0.3171
Oral cavity mean dose (β_2)[b]	0.1879
Age (β_3)	0.0238
Intercept (β_0)	-4.5092

2.5 DCNN

Deep convolutional neural networks (DCNNs) are commonly used for pattern or object detection in video and image data. Most neural networks comprise of multiple layers performing different functions (*28*). Convolutional layers automatically extract features within the image by applying convolutional filters. The filter is moved over the image in every dimension, calculating the dot product of the filter elements with the corresponding image values in every step to create an activation map. Different filters are convolved with the input image to detect different features. Pooling layers are placed in between convolutional layers to reduce the spatial size of the network. This down sampling reduces the number of parameters as well as computational load and can help against overfitting. This is mostly done by average pooling or maximum pooling where either the average or the maximum value from a pool is saved while the others are dropped. (*29*) The last layers usually include a fully connected layer that is a complete connection between all individual activations from the layer before.

2.6 rCNN

Adding an indefinite number of layers to the previously described DCNN is not possible due to degradation issues as the accuracy gets saturated with increasing network depth. In a residual convolutional neural network (rCNN), layers are skipped when their effect on the model performance is zero or negative using skip connections. This keeps the network from decreased performance in deep layers (*30*). While this type of network was designed for very large numbers of layers it can work also for more shallow networks.

3. Materials and Methods

In the following the material and methods applied in this research will be described. This section will be split up in a general overview of the patient cohort, methods used for the fitting of the logistic regression NTCP model and developing the NTCP with an alternative tongue mucosa structure, and the methods used for training and analysing the deep learning prediction model.

3.1 Patient cohort

Data collection

Prospective data collection was performed for all patients irradiated for head and neck at the University Medical Center Groningen (UMCG) between 2007 and 2022, and the toxicity and clinical data was stored in the internal REDCAP system. The clinical data contained all available clinical variables as sex, age, tumour location and radiation strategy used, as well as patient and clinician scored toxicity records from during and after the treatment. For each patient in the cohort a planning CT was created prior to the treatment start. Clinicians of the UMCG delineated the target volume and neighbouring OAR on the planning CT, resulting in a set of structures combined in a structure file, referred to as RTSTRUCT. All RTSTRUCT files must at least contain the structures listed in Table 2 that are of interest to this and related projects. The dose plans were created in accordance with the Dutch treatment guidelines and the 3D dose volume file will in the following be referred to as RTDOSE. A correct match between the CT and RTSTRUCT, as well as between the CT and RTDOSE was checked using the Frame of Reference Unique Identifier (FoR UID) that must match for data within the same reference frame. For patients with missing structures, deep learning contouring (DLC) structures had to be used instead, where CT slices are used as the input to predict the RTSTRUCT. A visual check was performed on a few selected cranial and transversal CT slices per patient to determine bad artefacts created by metal fillings or implants.

Table 2: Overview of relevant structures that must be part of the available RTSTRUCT set. "External" referrs to the entire region of the body included in the CT and therefore the radiation therapy planning.

Buccal mucosa left	Parotid gland left
Buccal mucosa right	Parotid gland right
Cervical oesophagus	Pharyngeal constrictor muscle (PCM) inferior
Cricopharyngeal muscle (Crico)	Pharyngeal constrictor muscle (PCM) medior
External	Pharyngeal constrictor muscle (PCM) superior
Oral cavity	Submandibular gland left
Supraglottic larynx area	Submandibular gland right

Treatment

All patients were treated according to the Dutch guidelines for head and neck cancer and in this case more specifically squamous cell carcinoma. Patients with stage I or II received accelerated radiotherapy under the age of 70 and conventional fractionated radiotherapy above the age of 70. For patients with advanced stages III and IV the treatment guideline advises conventional fractionated radiotherapy for patients above the age of 70, while younger patients were treated with concurrent platinum based chemoradiation. In case chemotherapy was not possible due to the general fitness of the patient, accelerated radiotherapy with or without weekly cetuximab was given. For conventional fractionated radiotherapy 33 or 35 fractions of 2.0 Gray were given five times per week and for accelerated radiotherapy 33 or 35 fractions of 2.0 Gray were given six times per week.

Eligibility criteria

The inclusion criteria for patients used in this project include a minimum age of 18 years at the start of the radiotherapy treatment, squamous cell carcinoma (SCC) tumour histology, no previous surgery of the tumour or a neck dissection prior to the radiotherapy treatment. An exception was made for tonsillectomy. Furthermore, no previous irradiation of the head and neck region can be reported, patients have a M0 status and therefore no metastasis and they must have received treatment with curative intent. A complete-case analysis was chosen where only patients with available endpoint and baseline data were considered in the analysis. Week 1 was being used as baseline due to the multitude of missing week 0 scores.

Endpoints

The taste loss score is recorded based on a 4-point Likert scale with the patient reported items none ("helemaal niet"), mild ("een beetje"), moderate ("nogal") and severe ("heel erg") in the EORTC QLQ-H&N35 questionnaire. These scores are recorded weekly during the treatment, as well as every 6 months after the treatment for the first two years and afterwards yearly as part of the follow up procedure. In this project none-mild and moderate-severe were combined to allow for comparability with the Van den Bosch et al. study.

Data organization

The CT, RTSTRUCT and RTDOSE files were retrieved for all eligible patients. Mean dose and volume measures for all OAR were calculated with a MATLAB (version MATLAB R2018a) script using the available CT, RTSTRUCT and RTDOSE files. For structures located on two sides, like the parotid gland, combined structures were defined, and their mean dose and volume parameters calculated in addition to the individual single sides. Dose-volume histograms (DVHs), which are histograms relating radiation dose to tissue volume, were the base of the mean dose calculations.

The data set was split into a training (70%), validation (15%), and test set (15%), by choosing a random seed under the condition of comparable event rates in all subsets. An appropriate seed is important to be able to reproduce network performance when training a neural network and ensure a comparable patient cohort in the training, validation, and test sets. The seed is the initialization state of a pseudo-random number generator determining the distribution of data into specified subsets. This means that the same seed used twice will generate the exact same random distribution of data in both cases.

Performance measures

All model performances were primarily evaluated and compared using the area under the curve (AUC). A higher AUC means that the model performs better at distinguishing between the positive and negative classes, so in this case with or without taste loss at 6 months. Additionally, R^2 tests, Hosmer–Lemeshow χ^2 tests and the calibration and discrimination slopes were used as a measure of good fit. Calibration curves were also plotted to allow for an additional visual comparison.

3.2 Reference NTCP model for taste loss

The NTCP model for taste loss at 6 months, as detailed in the paper by Van den Bosch et al (*17, 27*), served as baseline model in the current study. The age at the start of the radiation therapy treatment as well as the mean dose to the parotid glands and the oral cavity were included in the logistic regression model according to Formula (2). Note that the logarithm of the parotid gland mean dose and the square root of the oral cavity mean dose were used.

$$p = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}} \quad (2)$$

With:

$$\beta_0 = -4.5092$$

$$\beta_1 x_1 = 0.3171 \times \ln(Parotid\ mean\ dose)$$

$$\beta_2 x_2 = 0.1870 \times \sqrt{Oral\ cavity\ mean\ dose}$$

$$\beta_3 x_3 = 0.0238 \times Age$$

A performance evaluation on our patient cohort was performed using the coefficients of the Van den Bosch et al. model. For the training patient cohort, a refit of the reference model was created by calculating new predictor coefficients (β_1 – β_3 coefficient) as well as a new intercept (β_0 coefficient). This represents the best possible fit in the training population used in this project. These coefficients were then applied to create a logistic regression model for the validation and test sets.

3.3 NTCP model with tongue mucosa structure

Stieb et al. (*18*) introduced a contouring guideline for the taste bud bearing tongue mucosa with the intention to allow for more precise analysis of the dose distribution within the oral cavity. In the available CT images the whole tongue as well as the taste bud bearing tongue mucosa were delineated using two different methods. The difference between the whole tongue delineation and the commonly used oral cavity contour (*31*) is an exclusion of the soft palate, uvula and air pockets that can exist due to the positioning of the tongue. Both methods for the contouring of the taste bud region aim to cover the upper surface of the tongue. Method A does this using an adaption/subtraction method with the whole tongue structure as a starting point, while method B uses a freehand approach. A thickness of 5 mm was used with both methods to make dosimetric analysis as well as potential imaging biomarker evaluation possible.

This contouring guideline was approximated to evaluate the performance of an NTCP model using the mean dose to the taste bud bearing tongue mucosa instead of the entire oral cavity mean dose. However, since a manual delineation of the tongue mucosa structure in the entire patient cohort was not feasible for this project, it was derived from the oral cavity structure in MATLAB.

All patients were divided in three subsets based on visual inspection of their CT images. The first subset contained all patients that showed strong metal artefacts in the region of the oral cavity, due to dental fillings or implants. The second subset contained patients that were either treated with an open mouth or had large air-filled gaps between the tongue and the palate caused by an unusual placement of the tongue. All other patients that did not fall into either of these two subsets were labelled “normal”. Depending on the subset a slightly different technique was used to create the structure of the taste bud bearing tongue mucosa, i.e., tongue mucosa structure. The techniques chosen aimed to make a conservative structure that includes the upper layer of the tongue while excluding the palate in all patients.

Using the oral cavity structure as a starting point, the most cranial layer of voxels was removed for normal and artefact subsets and the three most cranial voxel layers were removed for the air gap subset. Air was filtered out by applying a Hounsfield thresholding with a lower cut-off at -500 HU and an upper at 500 HU. From the resulting tongue structure, the most cranial layer of four voxels thickness was selected as the taste bud bearing tongue mucosa for normal and air gap subsets. This equated to a thickness of 8 mm. For the artefact subset the most cranial layer of 8 voxels was selected with a thickness of 16 mm. The posterior quarter of the structure was then cut off to resemble the

length indicated by Stieb et al. (Figure 1). The corresponding mean dose and volume were calculated using the DVH.

The tongue mucosa mean dose was compared to the originally used oral cavity mean dose, as well as the parotid gland mean dose and age in a univariable analysis. Furthermore, a refit of the reference model was performed using the tongue mucosa mean dose instead of the oral cavity mean dose to compare their performance.

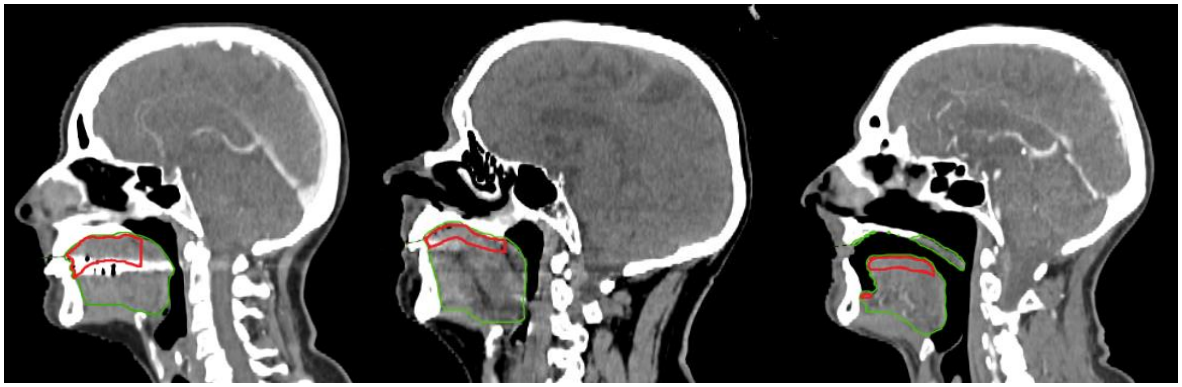

Figure 1: The three different sub-groups of patients with strong metal artefacts (left), normal tongue position (middle) and large airgap or open mouth (right) with the original oral cavity structure shown in green and the computed tongue mucosa structure in red.

3.4 Deep learning model

Pre-processing for deep learning model

For every patient where the FoR UID between the CT, RTDOSE and RTSTRUCT files matched, the dose and CT were registered to ensure a perfect overlay and saved as arrays. The relevant structures were extracted from the available RTSTRUCT files, and the resulting segmentation map was saved to a separate folder. A cropping region was defined to reduce all files to the same spatial dimensions by use of a bounding box. For the three spatial dimensions, the conditions for the centre of the bounding box placement were as follows:

- Z centre: centre coordinate between the most upper voxel (with value > 0) in z-dimension in the segmentation map of the parotid glands and the lowest voxel (with value > 0) in z-dimension in the segmentation map of the crico, thyroid and mandible.
- Y centre: centre coordinate between the most upper and lowest voxel (with value > 0) in y-dimension in the segmentation map of the parotid glands.
- X centre: centre coordinate between the most upper and lowest voxel (with value > 0) in x-dimension in the segmentation maps of the parotid and submandibular glands.

The bounding box for this project was chosen to have the dimensions of 100 x 100 x 100 voxels at a spacing of 2 x 2 x 2 mm^3 and the cropped arrays therefore had the same size. Other operations that were done if needed included spacing correction, value clipping in the CT images and transformations like resizing on all three components (CT, RTDOSE, segmentation map). Since the resulting arrays all had the same dimensions, they could be concatenated into a single NumPy array.

Patients with a taste toxicity score of “None” and “Mild” at 6 months were appointed label “0” while patients with a taste toxicity score of “Moderate” and “Severe” were appointed label “1”. For each patient a set of 2D subplots of the CT, RTDOSE and segmentation map was created in order to be able to do visual checks for irregularities (Figure 2).

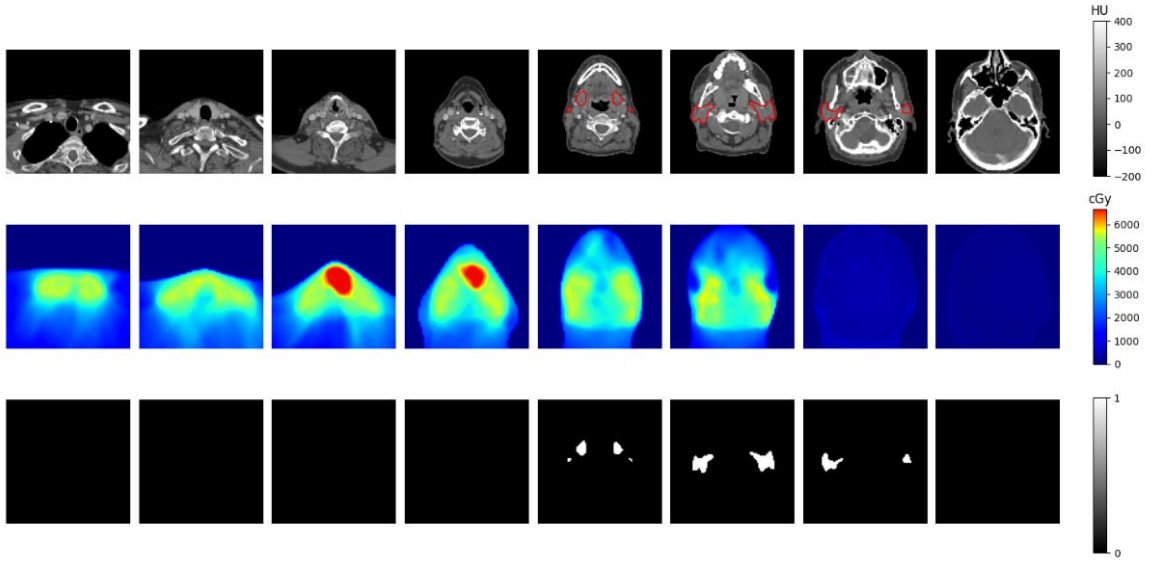

Figure 2: Overview plot created of the CT (top), dose (middle) and segmentation map (bottom) during the preprocessing step.

Systematic training of deep learning model

Multiple deep learning architectures were trained with systematically adjusted parameters. Train-validation-test experiments were conducted to evaluate the performance of the trained networks with an initial focus on the AUC values of the validation and test set.

The preferred network architectures were chosen based on short test runs on a subset of the patient cohort as well as literature (*26*) and a DCNN, as well as a rCNN were used. The order in which network training parameters were adjusted systematically was based on their degree of impact on a network. The default starting parameters used for both architectures can be seen in Table A2 (Appendix C), and all adjustments described in the following were done by only adjusting the parameter of interest while leaving all previously determined parameters the same.

In an initial step different learning rate schedulers for the cosine scheduler-function were tested between the range of $T_0 = 2$ and $T_0 = 40$, where T_0 refers to the number of epochs before starting a new scheduler iteration. This is demonstrated in Figure 3. The RMSprop and stochastic gradient descent (SGD) optimizer functions were tested in combination with this, and both are versions or adaptations of the commonly used gradient descent technique. They aim to minimize the loss function in steps whose size is defined by a learning rate. They were picked based on previous success in a related project on predicting xerostomia in the same patient cohort within the research group. The RMSprop optimizer can increase the model's learning rate by restricting oscillations perpendicular to the gradient direction. It achieves this by dividing the learning rate by the moving average of squared gradients (*32*). The stochastic gradient descent (SGD) is an optimization of the traditional gradient descent where instead of all data points in the training set a single data point is used to calculate the updated model parameter, resulting in faster iterations (*32*).

Batch sizes between batch = 2 and batch = 32 were tested; the batch size defines the number of data samples that are being passed through the neural network at the same time. Loss functions are used to measure the model's prediction performance and it must be minimized as the model is being optimized. Four different loss functions were tested: cross-entropy, F1, L1 and ranking. The cross-entropy loss function facilitates wrong predictions to be penalized according to a logarithmic scale, meaning that larger differences between the predicted and the true value are penalized more than

small differences. Furthermore, the F1 loss that minimizes the harmonic mean and the L1 loss function which minimizes the absolute error, or the sum of all the absolute differences between true and predicted values were tested. The ranking loss function utilizes metric learning to calculate differences between images.

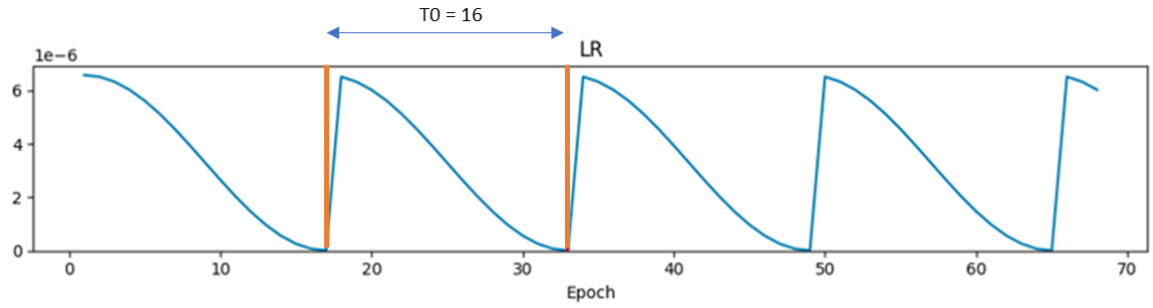

Figure 3: Demonstration of a cosine learning rate finder function with T_0=16. The orange lines show the start and end of one cycle that is iterated after T_0 epochs.

After each set of tests, the best performing networks per architecture were chosen based on their AUC values and included in the following test. Networks that reached their best performance at less than 5 epochs were excluded, as well as networks showing clear signs of overfitting. As a final step age was added as an additional predictor in the last layer of both best performing networks to determine the added value.

In addition to the previously described performance measures, attention maps were created using the Grad-CAM++ method by Chattopadhyay (*33*). These attention maps are human-interpretable visual explanations of the classification task where areas the neural network pays more and less attention to are made visible.

4. Results

4.1 Patient cohort

Many patients had not filled in all follow up questionnaires and 422 were excluded due to missing endpoint data at 1 week and 6 months. After all exclusion steps 949 patients were included to be used as part of this project (Appendix B, Figure A1; Table 3). There were no significant differences between the subsets for training, validation and testing regarding clinical parameters and average mean dose values. For 16 patients DLC structures were used instead of pre-existing RTSTRUCTS due to missing structures. An overview of relevant OAR dose parameters with the average mean dose calculated on the entire cohort and all sub-cohorts and their standard deviation can be seen in Table 4.

The number of included patients treated yearly in the UMCG for HNC increased from 40 (33 male and 7 female) in 2007 to 84 (64 male and 20 female) in 2020 (Figure 4). 76 patients treated in the second half of 2021 and the beginning of 2022 had not completed the follow up (FU) trajectory enough to be able to be included for this project. The distribution of endpoint data in this patient cohort showed 26.5 % of patients experiencing moderate – severe taste loss at 6 months compared to 4.3 % at week 1 (Figure 5).

Table 3: Patient, tumour and treatment characteristics for the entire patient cohort in comparison to the training, validation and test sub-cohorts.

	All		Training		Validation		Test		p-value
n	949		664		143		142		
Sex (%)									0.32
Male	711	(74.9)	494	(74.4)	114	(79.7)	103	(72.5)	
Female	238	(25.1)	170	(25.6)	29	(20.3)	39	(27.5)	
Age (SD)	64	(9.8)	63.6	(10.0)	64.3	(9.0)	64.5	(10.8)	0.56
Technique (%)									0.61
3D-CRT	69	(7.3)	42	(6.3)	14	(9.8)	13	(9.2)	
IMRT	426	(44.9)	303	(45.6)	59	(41.3)	64	(45.1)	
IMPT	133	(14)	90	(13.6)	24	(16.8)	19	(13.4)	
VMAT	321	(33.8)	229	(34.5)	46	(32.2)	46	(32.4)	
Loctum (%)									0.53
Other	4	(0.4)	3	(0.5)	1	(0.7)	0	(0)	
Hypopharynx	79	(8.3)	57	(8.6)	11	(7.7)	11	(7.7)	
Larynx	416	(43.8)	294	(44.3)	55	(38.5)	67	(47.2)	
Nasopharynx	38	(4)	26	(3.9)	8	(5.6)	4	(2.8)	
Nasal cavity	7	(0.7)	4	(0.6)	0	(0)	3	(2.1)	
Oral Cavity	46	(4.8)	33	(5)	9	(6.3)	4	(2.8)	
Oropharynx	359	(37.8)	247	(37.2)	59	(41.3)	53	(37.3)	
P16 (%)									0.75
Negative	217	(22.9)	150	(22.6)	37	(25.9)	30	(21.1)	
Not determined	563	(59.3)	400	(60.2)	78	(54.5)	85	(59.9)	
Positive	169	(17.8)	114	(17.2)	28	(19.6)	27	(19)	
Modality (%)									0.52
Accelerated RT	306	(32.2)	214	(32.2)	45	(31.5)	47	(33.1)	
Cetuximab	40	(4.2)	34	(5.1)	3	(2.1)	3	(2.1)	
Chemoradiation	343	(36.1)	238	(35.8)	51	(35.7)	54	(38)	
Conventional RT	260	(27.4)	178	(26.8)	44	(30.8)	38	(26.8)	
Taste loss week 1 (%)									0.9
None	768	(80.9)	542	(81.6)	110	(76.9)	116	(81.7)	
Mild	141	(14.9)	95	(14.3)	26	(18.2)	20	(14.1)	
Moderate	29	(3.1)	19	(2.9)	5	(3.5)	5	(3.5)	
Severe	11	(1.2)	8	(1.2)	2	(1.4)	1	(0.7)	
Taste loss 6 months (%)									0.61
None	442	(46.6)	322	(48.5)	57	(39.9)	63	(44.4)	
Mild	255	(26.9)	173	(26.1)	43	(30.1)	39	(27.5)	
Moderate	170	(17.9)	112	(16.9)	31	(21.7)	27	(19)	
Severe	82	(8.6)	57	(8.6)	12	(8.4)	13	(9.2)	
T Stage (%)									0.87
Tis – T2	455	(47.9)	324	(48.8)	64	(44.8)	67	(47.2)	
T3 – T4	493	(51.9)	339	(51.1)	79	(55.2)	75	(52.8)	
Tx	1	(0.1)	1	(0.2)	0	(0)	0	(0)	
N Stage (%)									0.93
N0	445	(46.9)	312	(47)	64	(44.8)	69	(48.6)	
N+	503	(53)	351	(52.9)	79	(55.2)	73	(51.4)	
Nx	1	(0.1)	1	(0.2)	0	(0)	0	(0)	
Neck irradiation (%)									0.19
No	191	(20.1)	141	(21.2)	20	(14)	30	(21.1)	
Bilateral	724	(76.3)	502	(75.6)	118	(82.5)	104	(73.2)	
Unilateral	34	(3.6)	21	(3.2)	5	(3.5)	8	(5.6)	
Contrast (%)	559	(59)	384	(58)	80	(56)	95	(67)	
Artefact (%)	199	(21)	134	(20)	29	(20)	36	(25)	

Table 4: Dose parameters for the entire patient cohort in comparison to the training, validation and test sub-cohorts.

Structure	All		Training		Validation		Test		p-value
Parotid (SD)									
Parotid left	21.2	(15.5)	20.8	(15.6)	22.7	(15.1)	21.4	(15.8)	0.40
Parotid right	22.3	(15.8)	21.9	(15.7)	24.5	(16.1)	21.6	(16.2)	0.18
Combined	21.7	(14.7)	21.4	(14.7)	23.6	(14.1)	21.6	(15.2)	0.26
Submandibular (SD)									
Submandibular left	44.8	(23.0)	44.4	(23.4)	47.66	(21.1)	44.2	(23.0)	0.29
Submandibular right	45.7	(23.3)	45.4	(23.7)	48.6	(21.6)	44.6	(23.2)	0.27
Combined	45.4	(22.0)	45.0	(22.5)	48.2	(19.7)	44.8	(21.8)	0.26
PCM (SD)									
PCM medior	44.5	(20.5)	43.9	(21.0)	46.5	(19.4)	45.6	(19.2)	0.29
PCM superior	39.2	(23.2)	38.6	(23.4)	42.8	(21.0)	38.4	(23.8)	0.13
PCM inferior	48.4	(17.0)	47.8	(17.4)	48.9	(16.2)	50.5	(15.5)	0.22
Combined	43.2	(16.0)	42.6	(16.4)	45.9	(13.7)	43.2	(16.1)	0.08
Crico (SD)	40.1	(16.3)	39.5	(16.7)	40.9	(14.9)	42.1	(15.3)	0.20
Supraglottic (SD)	50.2	(16.3)	49.8	(16.5)	50.6	(16.7)	51.7	(14.4)	0.44
Oral Cavity (SD)	31.7	(21.4)	31.2	(21.5)	34.6	(20.6)	31.4	(21.8)	0.22
Buccal mucosa (SD)									
Buccal mucosa left	25.3	(21.1)	24.8	(21.0)	28.2	(20.8)	25.0	(22.0)	0.23
Buccal mucosa right	26.4	(21.9)	25.9	(21.9)	29.7	(22.2)	25.5	(21.7)	0.15
Combined	26.0	(20.1)	25.4	(19.9)	29.1	(19.9)	25.4	(20.6)	0.14
Oesophagus (SD)	28.2	(18.7)	27.5	(18.9)	29.7	(17.6)	29.6	(18.4)	0.26
External (SD)	10.1	(5.8)	10.0	(5.9)	10.5	(5.3)	10.3	(6.2)	0.66

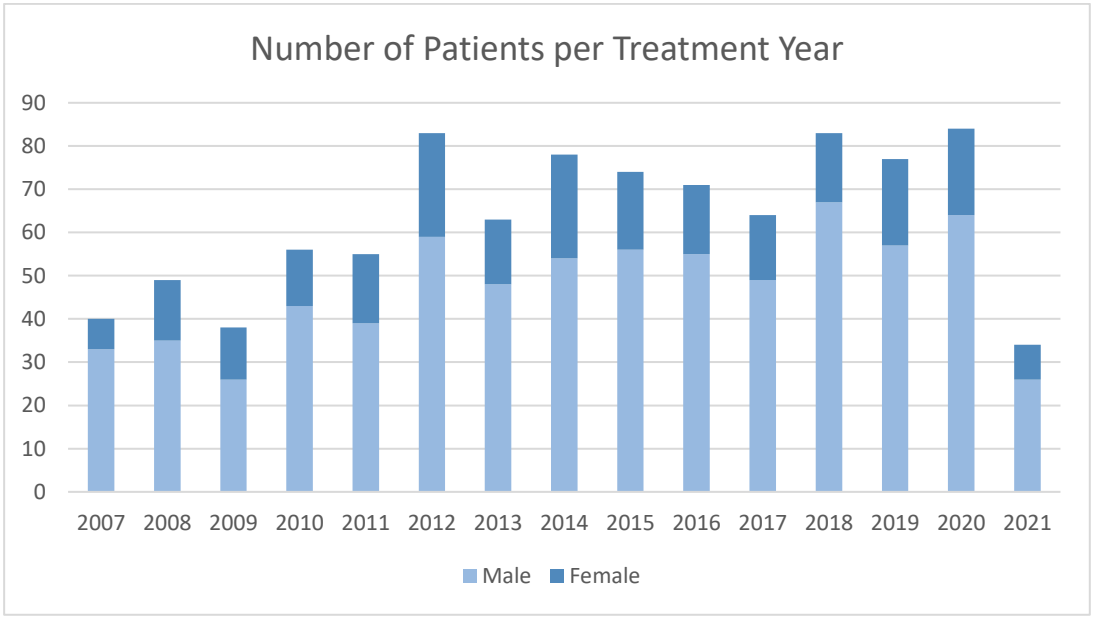

Figure 4: Number of patients treated with radiation therapy for HNC at the UMCG each year.

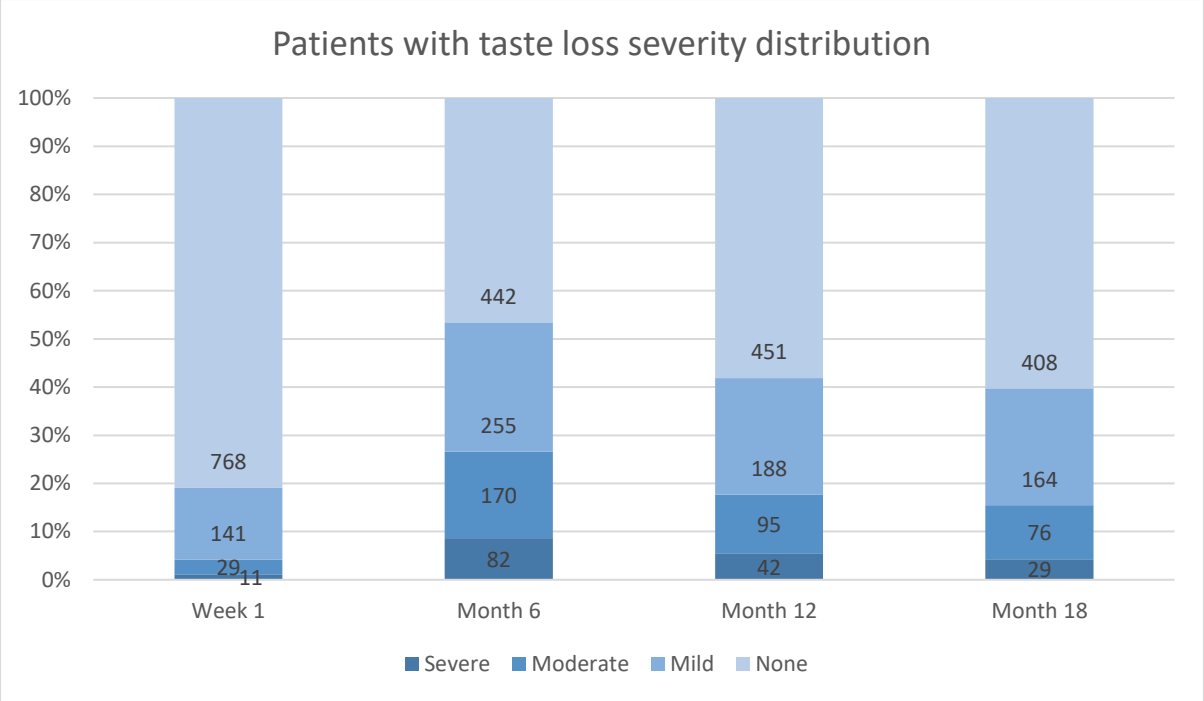

Figure 5: Distribution of endpoint data of this cohort in percent, with most patients experiencing taset loss symptoms at 6 months.

4.2 Baseline NTCP model

The results of the performance evaluation using the coefficients from the NTCP of Van den Bosch et al. model on the training set yielded an AUC of 0.723 (95% CI: 0.682 – 0.766) and a calibration slope of 1.17 (Table 5, Figure 6). The refit performed on the training set led to an increased AUC of 0.724 (95% CI: 0.683 – 0.766) and a calibration slope of 1.01. The refit coefficients obtained from the refit of the training set (Table 6) applied to the validation and test subsets resulted in a logistic regression model with AUC performances of 0.706 (95% CI: 0.621 – 0.791) and 0.692 (95% CI: 0.595 – 0.788) and calibration slopes of 1.11 and 0.88 respectively.

Table 5: The reference model performance on the training set and the refit performance on the training, validation and test subsets.

	Performance evaluation Van den Bosch et al.	**Refit**		
	Training Set	Training Set	Validation Set	Test Set
R^2	0.114	0.117	0.121	0.097
AUC (95% CI)	0.723 (0.682 – 0.764)	0.724 (0.683 – 0.766)	0.706 (0.621 – 0.791)	0.692 (0.595 – 0.788)
HL test χ^2 (p-value)	3.93 (0.86)	5.13 (0.74)	5.83 (0.67)	11.69 (0.17)
Calibration slope (intercept)	1.17 (-0.04)	1.00 (-0.00)	1.11 (-0.02)	0.88 (0.05)
Discrimination slope	0.097	0.117	0.102	0.103

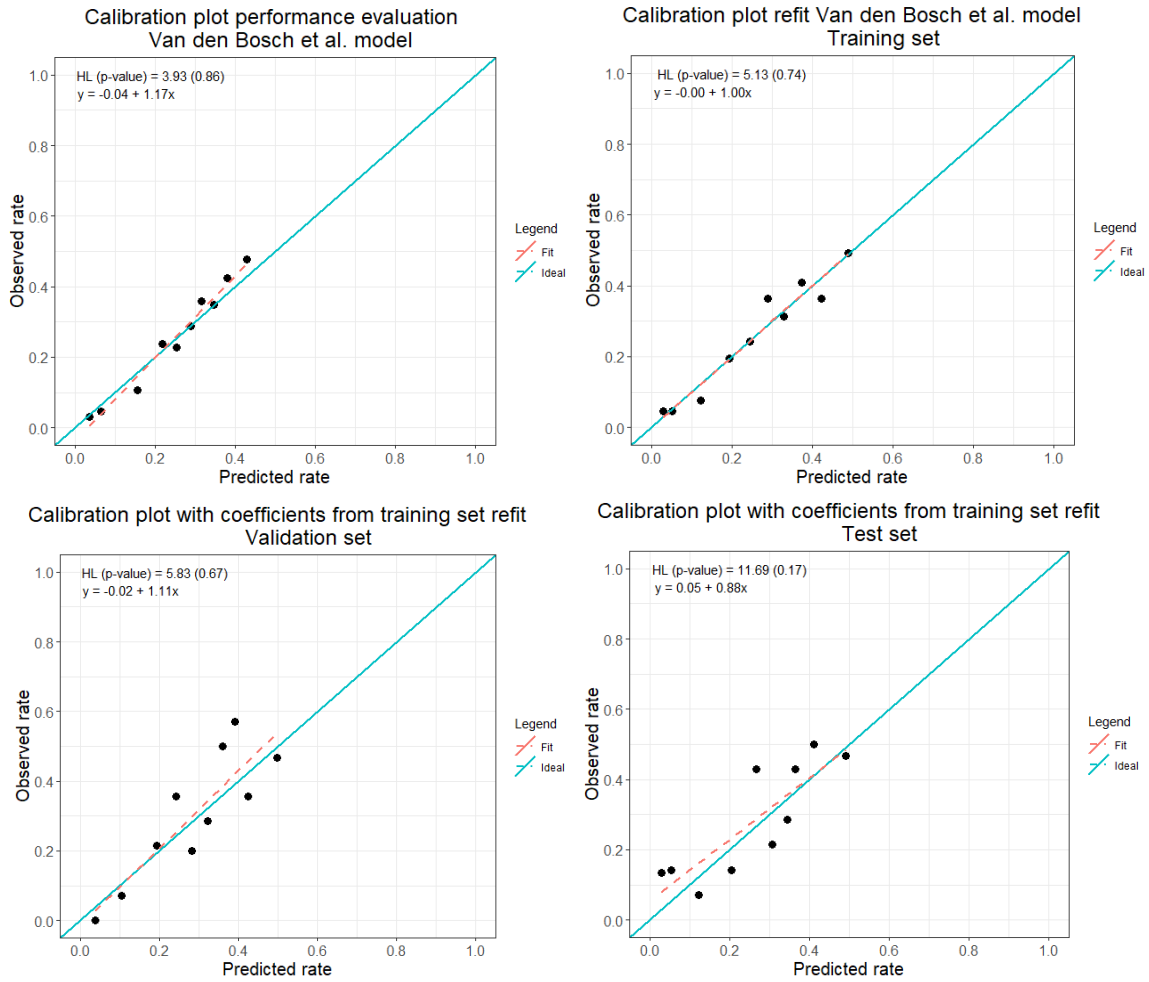

Figure 6: Calibration plots of the logistic regression models. Top left: Plot of the performance evaluation using the Van den Bosch et al. coefficients on the training set. Top right: Plot of the refit performed on the training set resulting in a new set of coefficients. Bottom left: Plot of the refit model applied to the validation set. Bottom right: Plot of the refit model appplied to the test set.

Table 6: Coefficients of the original Van den Bosch et al. model and the refit coefficients used for the baseline model of this patient cohort. (a: ln(mean dose to both parotid glands); b: sqrt(mean dose to oral cavity))

Model Predictors	Van den Bosch et al. model coefficients	Refit model coefficients
Intercept	-4.5092	-5.4862
Parotid glands [a]	0.3171	0.2042
Oral cavity [b]	0.1870	0.3272
Age	0.0238	0.0318

4.3 NTCP model with tongue mucosa structure

There was a strong correlation (Pearson coefficient = 0.94, p = 2.2 x10^{-16}) between the tongue mucosa mean dose and the oral cavity mean dose (Appendix B, Figure A2) with on average lower mean dose values in the tongue mucosa than in the oral cavity. The univariable analysis showed best performance for the oral cavity mean dose in terms of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) followed by the tongue mucosa mean dose, parotid mean dose and age (Table 7). When comparing oral cavity mean dose and tongue mucosa mean dose, the oral cavity mean dose had superior performance across all measures. It showed a higher AUC of 0.697 (Tongue mucosa: 0.690); notably, the oral cavity had a larger coefficient of 0.016 (Tongue mucosa: 0.013), which was

met with a higher odds ratio of 1.037 (Tongue mucosa: 1.030), indicating that the oral cavity mean dose has a larger influence on the prediction of the logistic regression model. The age parameter was not significant in the univariable analysis and had little effect on the logistic regression model with an odds-ratio of 1.006 (0.991 – 1.021).

The refit of the logistic regression model performed on the training set with the oral cavity mean dose yielded an AUC of 0.724 and calibration slope of 1.01 (Table 8). When refit with the tongue mucosa mean dose, the logistic regression model had an AUC of 0.717 and calibration slope of 1.02. All logistic regression coefficients changed as a result of exchanging the oral cavity mean dose with the tongue mucosa mean dose (Table 9). A larger coefficient of 0.327 was observed for the oral cavity mean dose compared to the tongue mucosa mean dose coefficient of 0.188. The similarity in calibration slope and intercept can be seen in Figure 7.

Table 7: Univariable analysis results sorted by p-value.

	P - value	AUC	Nagelkerke R^2	OR (95% CI)	AIC	BIC	Coefficient
Oral Cavity mean dose	<0.0001	0.697	0.137	1.037 (1.029-1.045)	1008.9	1018.7	0.016
Tongue mucosa mean dose	<0.0001	0.690	0.107	1.030 (1.023 – 1.038)	1030.4	1040.2	0.013
Parotid mean dose	<0.0001	0.669	0.091	1.041 (1.030 – 1.053)	1041.2	1050.9	0.017
Age	0.428	0.514	0.001	1.006 (0.991 -1.021)	1101.9	1111.6	0.003

Table 8: Logistic regression model performance with the oral cavity mean dose and tongue mucosa mean dose in comparison.

	Refit with oral cavity mean dose	Refit with tongue mucosa mean dose
R^2	0.117	0.109
AUC (95% CI)	0.724 (0.683-0.766)	0.717 (0.676 – 0.758)
HL test χ^2 (p-value)	5.14 (0.74)	2.54 (0.96)
Calibration slope (intercept)	1.01 (-0.00)	1.02 (-0.00)
Discrimination slope	0.117	0.110

Table 9: Logistic regression coefficients resulting from the refit with the oral cavity mean dose and the tongue mucosa mean dose in comparison. (a: ln(mean dose to both parotid glands); b: sqrt(mean dose to oral cavity); c: sqrt(mean dose to tongue mucosa))

Model Predictors	Refit coefficients with oral cavity mean dose	Refit coefficients with tongue mucosa mean dose
Intercept	-5.4862	-4.9554
Parotid glands [a]	0.2042	0.4052
Oral cavity [b]	0.3272	
Tongue mucosa [c]		0.1876
Age	0.0318	0.0293

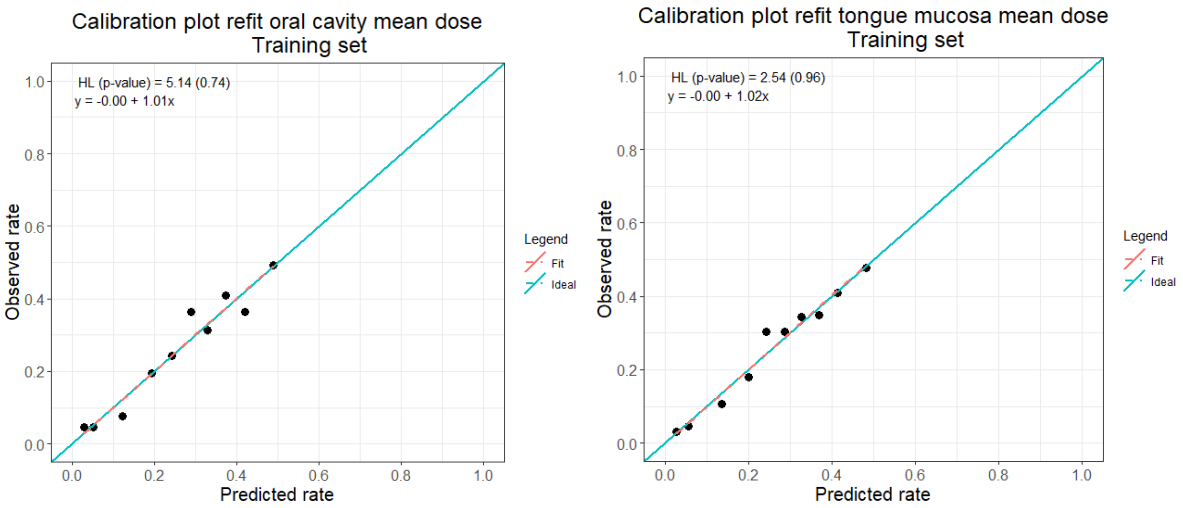

Figure 7: Calibration plots of the logistic regression models using the oral cavity mean dose (left) and the tongue mucosa mean dose (right) to perform the refit.

4.4 Deep learning model

The systematic adjustment of network parameters seen in Tables 10 to 12 lead to one best performing network for each of the two architectures used. The best performing DCNN yielded a validation AUC of 0.719 and the test on an independent subset from the same overall patient cohort gave an AUC of 0.682 (Table 13). It used a factor three higher label weights on the taste loss group during training in comparison to the no taste loss group. Further details about the model can be found in appendix C. The best performing rCNN had a validation AUC of 0.698 and a test AUC of 0.684 (Table 13). The addition of age in the last layer of the networks led to overfitting issues in both cases and was therefore removed again.

Additionally, the calibration plots can be seen in Figure 8 with calibration slopes of 0.71 and 0.63 for the DCNN and rCNN respectively. Both calibration and discrimination slopes of the neural networks show worse performance than the logistic regression baseline model.

Table 10: Systematic adjustment of network parameters: Testing the optimizer functions RMSProp and SGD with different learning rate scheduler times between $T_0 = 8$ and $T_0 = 40$. For every combination the AUCs are given for the training, validation and test set, as well as the number of epochs at which the maximum performance was reached. Red markings indicate failed training, yellow markings indicate overfitting tendencies and green boxes indicate the best performing networks that were taken to the next round of systematic parameter adjustment.

		$T_0 = 8$			**$T_0 = 16$**			**$T_0 = 32$**			**$T_0 = 40$**		
	AUC	Train.	Val.	Test	Train.	Val.	Test	Train.	Val.	Test	Train.	Val	Test
DCNN	RMSProp	0.963	0.731	0.64	**0.875**	**0.719**	**0.682**	0.885	0.717	0.676	0.883	0.723	0.668
		Epoch 325			**Epoch 132**			Epoch 139			Epoch 139		
	SGD	0.828	0.703	0.682	0.804	0.689	0.686	0.807	0.685	0.714	0.943	0.711	0.659
		Epoch 18			Epoch 6			Epoch 2			Epoch 93		
rCNN	RMSProp	0.862	0.661	0.617	0.867	0.687	0.665	0.866	0.667	0.628	0.907	0.682	0.647
		Epoch 1			Epoch 1			Epoch 1			Epoch 2		
	SGD	0.789	0.693	0.675	0.776	0.683	0.679	0.739	0.69	0.666	**0.744**	**0.699**	**0.66**
		Epoch 222			Epoch 188			Epoch 108			**Epoch 96**		

Table 11: Systematic adjustment of network parameters: Testing batch sizes between Batch = 4 and Batch = 32. For every combination the AUCs are given for the training, validation and test set, as well as the number of epochs at which the maximum performance was reached. Red markings indicate failed training, yellow markings indicate overfitting tendencies and green boxes indicate the best performing networks that were taken to the next round of systematic parameter adjustment.

	Batch = 4			**Batch = 8**			**Batch = 16**			**Batch = 32**		
	Train.	Val.	Test	Train.	Val.	Test	Train.	Val.	Test	Train.	Val.	Test
DCNN, T0 = 16	0.902	0.736	0.68	**0.875**	**0.719**	**0.682**	0.919	0.703	0.67	0.912	0.702	0.669
	Epoch 25			**Epoch 132**			Epoch 117			Epoch 67		
rCNN, T0 = 40	0.782	0.619	0.654	**0.744**	**0.699**	**0.66**	0.724	0.687	0.63	0.668	0.596	0.524
	Epoch 4			**Epoch 96**			Epoch 41			Epoch 128		

Table 12: Systematic adjustment of network parameters: Testing the loss functions CE, F1, L1 and Ranking. For every combination the AUCs are given for the training, validation and test set, as well as the number of epochs at which the maximum performance was reached. Red markings indicate failed training, yellow markings indicate overfitting tendencies and green boxes indicate the best performing networks that were used for the evaluation.

	CE			**F1**			**L1**			**Ranking**		
	Train.	Val.	Test	Train.	Val.	Test	Train.	Val.	Test	Train.	Val.	Test
DCNN, Batch 8	**0.875**	**0.719**	**0.682**	0.742	0.698	0.674	0.443	0.665	0.664	0.719	0.654	0.693
	Epoch 132			Epoch 2			Epoch 39			Epoch 34		
rCNN, Batch 8	0.744	0.699	0.66	**0.726**	**0.698**	**0.684**	0.626	0.626	0.581	0.636	0.66	0.623
	Epoch 96			**Epoch 42**			Epoch 131			Epoch 94		

Table 13: DCNN and RCNN model performance on the test set compared to the logistic regression baseline model performance on the test set.

	Baseline Test	**DCNN Test**	**RCNN Test**
R^2	0.097	0.078	0.086
AUC (95% CI)	0.692 (0.595 – 0.788)	0.682 (0.591 – 0.773)	0.684 (0.585 – 0.784)
HL test X^2 (p-value)	11.69 (0.17)	33.41 (5.21 x10^{-05})	46.21 (2.17 x10^{-07})
Calibration slope (intercept)	0.88 (0.05)	0.71 (-0.06)	0.63 (0.14)
Discrimination slope	0.103	0.114	0.141

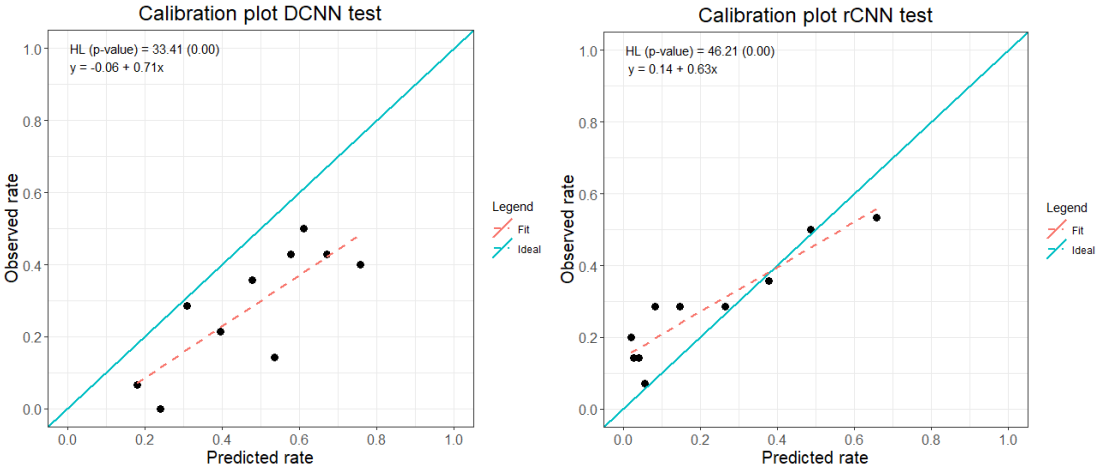

Figure 8: Calibration slopes of the DCNN model (left) and the rCNN model (right) when applied to the test set.

Visual analysis of the attention maps of the DCNN showed correlations between the highest dose depositions and the highest network attentions for 80% of patients in the test set. In contrast, only 25% of patients of the test set showed this strong correlation between dose and attention for the rCNN. The RTSTRUCTs seem to have a stronger influence for the rCNN with 75% of the patients showing strong correlation to either only the RTSTRUCTS or a mixture between dose and RTSTRUCT. A set of multiple attention maps for both networks can be seen in Appendix B (Figures A3 – A5).

5. Discussion

In this project, two methods were used to try and improve the NTCP model predicting late taste loss at six months post radiotherapy treatment by introducing a new structure of the tongue mucosa and using neural networks to create a new prediction model. While the performance of both methods was comparable to the original performance of the reference logistic regression model, neither performed better than the reference model.

Taste loss as a toxicity is not researched extensively and the exact reason for taste loss due to radiation therapy is complex. Main factors may be the disappearance of taste buds and dysfunction of the salivary glands (*14*). Saliva plays an important role in how we taste and perceive taste, making a connection between damage to salivary glands and side effects like xerostomia (dry mouth syndrome) likely. Saliva is responsible for the transport of taste substances to taste receptors and acts as a solvent for the taste substances in foods and drinks. It also acts as a protective layer to the taste receptors from damage by dryness or bacterial infection. A reduction in salivary secretion may therefore also decrease taste sensitivity (*34*). A smaller study has found potential reduction in symptom burden when treated with intensity modulated proton therapy (IMPT) instead of IMRT (*15*). Another source found no significance between oral cavity mean or maximum dose and the overall taste sensation in their cohort (*35*) in contrast to the predictors chosen by the Van den Bosch et al. model mentioned earlier. The number of comprehensive studies done on the influence of radiotherapy on taste loss and taste alteration is very limited, especially studies using larger patient cohorts.

The approach of replacing the entire oral cavity by a smaller structure that only encompasses the taste bud bearing tongue mucosa seemed promising but ended up not performing better in the logistic regression model predicting late taste toxicity at six months. The structures created were based on the technique described by Stieb et al. (*18*), but in this project the manual delineation was replaced by programmed estimation due to time constraints. In the future a smaller sub-cohort should be chosen for a more detailed comparison between the manual delineations and the derived structures used in this project. Furthermore, the focus of this method was placed on the taste buds and while they undoubtedly have an impact on a patient's ability to taste, their fast regeneration and turnover (*36*) might mean that this structure could yield better results in models for acute toxicity.

The deep learning approach has not been tested to its full potential yet. More combinations of parameters and architectures can be tried to improve the performance. The analysis of the attention maps allowed for a comparison of focus between the two neural network architectures used. While the DCNN showed mainly correlations with the high dose regions, the rCNN seemed more balanced and placed attention on low dose regions as well. This was the case especially in the RTSTRUCT regions like the parotid glands and oral cavity. A more in-depth analysis of the activation maps will give insight into what the focus of the neural network was more precisely and what a future version needs to improve upon.

One of the larger issues in this part of the project was the low event rate in all sub-cohorts with none of them containing more than 30% of patients with moderate-severe taste loss. When the network was trained without any adjustments or alterations, it would see less patients with symptoms and therefore adjusting hyperparameters based on the larger group of patients without symptoms. This can cause bias in the model as well as biased results. To counteract this effect, label weights were adjusted for the DCNN, giving more weight to data with an event. This could not be implemented for the rCNN as of yet but might improve the rCNN's performance in our patient cohort.

The comparability of the input data is important to assure that the network does not get distracted by irrelevant features in the data. In this case a problem might have been caused by the inconsistency of

CT images. The planning CT was used as an input for all patients, however 59% of the CTs used contained contrast, while the other 41% did not. 89% of patients predicted as true positives by both networks had a contrast CT as their planning CT (Appendix A, Table A1). On top of that, 100% of patients predicted as false positives by both networks had a contrast CT as their planning CT. Further testing of the network should be done with either only contrast CT images or only no contrast CT images, in order to be able to eliminate the possible bias introduced by using a mixture of both.

While a complete-case analysis was chosen for this project, this might be a disadvantage for our outcome. More data is better when training a neural network that should work as robustly as possible and by excluding patients where the endpoint was missing at six months or week one, 30% of the available data had to be excluded. Multiple imputation is a method often used to impute missing data based on available data points of the entire data set. This technique was used in the Van den Bosch et al. study and might be applicable in this case as well. However, this must be done with great caution as over or underestimation of the patients' symptoms must be avoided.

An additional analysis of association between taste loss and different predictors was done in different forms to improve our understanding of their relatedness. A correlation map containing clinical and treatment parameters, as well as OAR mean dose is shown in Figure 9. A strong correlation between the mean dose values could be seen in dark blue and a strong negative correlation between the mean dose values and the treatment year. The second was due to the improvements in OAR sparing with new radiotherapy techniques in recent years. However, no noticeably strong correlation between toxicity at week one or six months can be determined from this correlation matrix, stressing the importance of projects like this that try and improve the understanding of the mechanisms behind taste loss in HNC patients.

Based on the reference model predictors further analysis was done on the behaviour of oral cavity and parotid mean dose in this patient cohort. In the scatter plot (Figure 10) more red data points, representing patients experiencing late taste loss, can be seen towards the larger mean dose values for both parotid glands and oral cavity, showing the correlation between higher dose values in those structures and taste loss symptoms at 6 months. As Figure 11 demonstrates, there were higher mean dose values associated with patients experiencing taste loss at six months for both predictors. This effect is stronger for the oral cavity mean dose than for the parotid mean dose. In both parotid gland mean dose and oral cavity mean dose box plots that show the effect of higher mean doses in patients with late taste loss toxicity, a big gap can be seen between 'none' and 'mild'. This indicates that there might be an added benefit to evaluating only 'none' as no taste loss and all three categories of 'mild', 'moderate' and 'severe' as late taste loss.

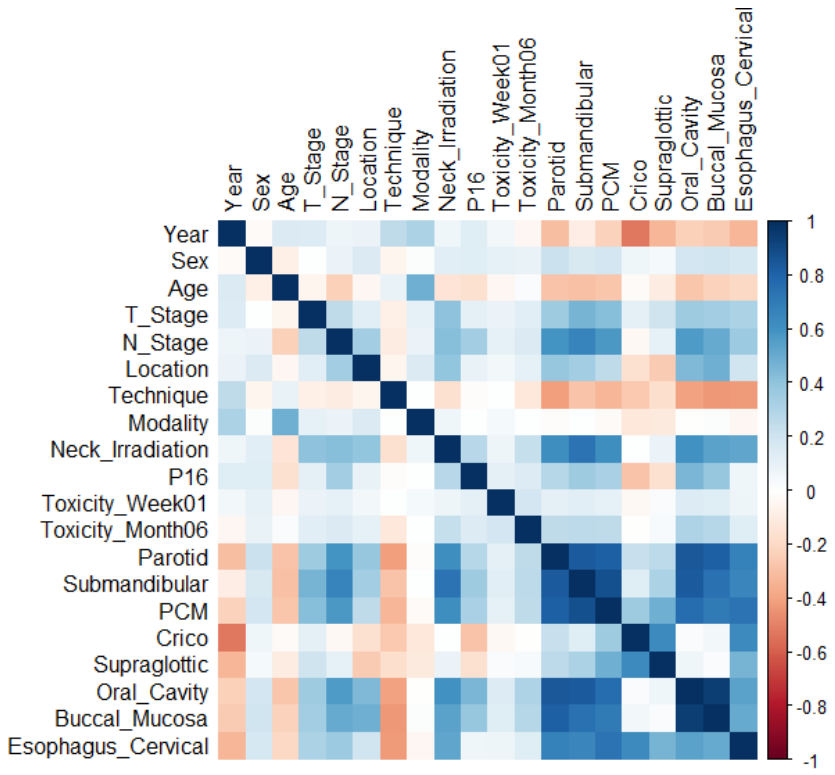

Figure 9: Correlation between clinical and dose parameters as well as taste loss (Toxicity week 1 and month 6) with dark blue representing a stong correlation and dark red representing a strong negative correlation. White represents no correlation.

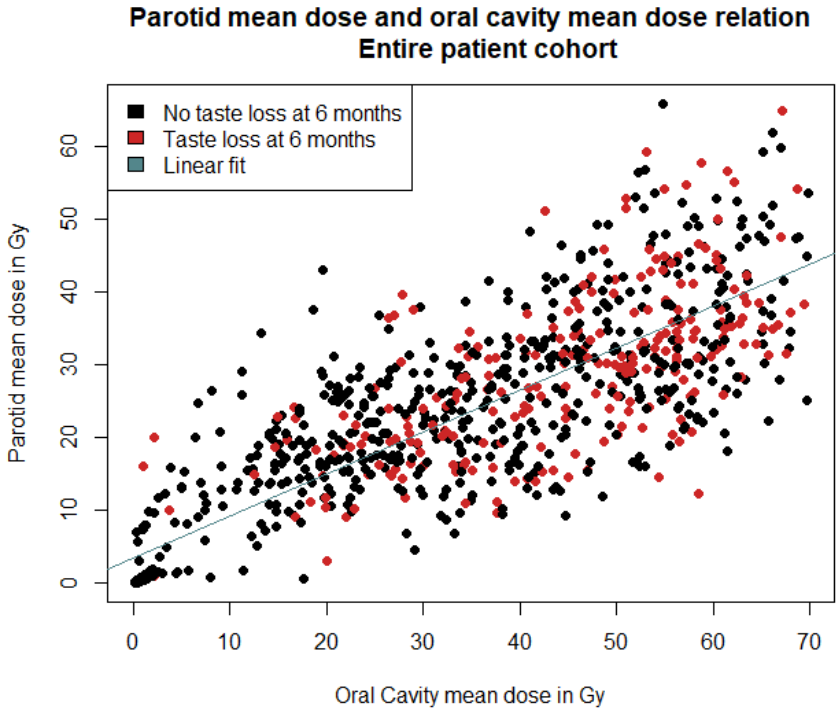

Figure 10: Scatter plot showing the relationship between parotid mean dose and oral cavity mean dose where patients experiencing taste loss at 6 months are shown in red. The linear fit with a sloe of 0.57 and an intercept of 3.50 is shown in blue.

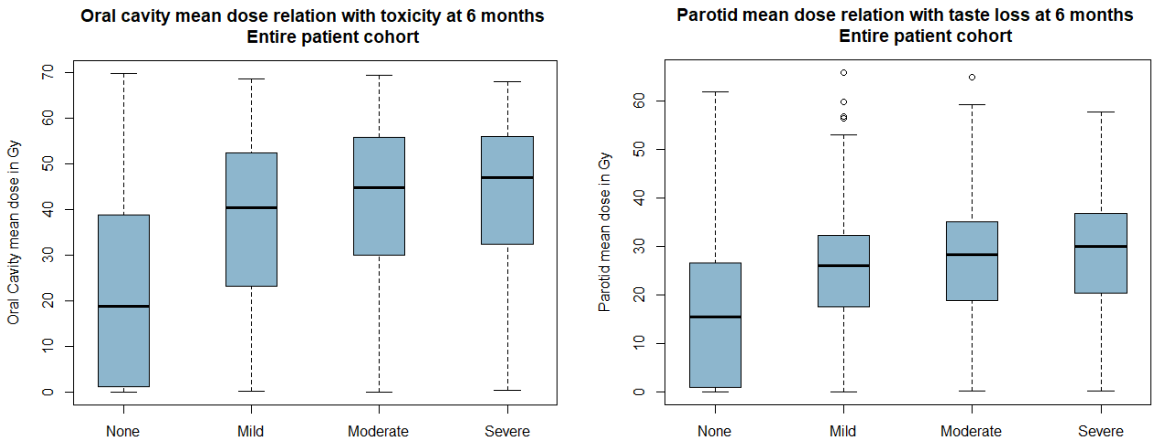

Figure 11: Box plots showing the average mean dose and the 25th – 75th percentile per toxicity score at 6 months post treatment for the oral cavity (left) and the parotid glands (right).

6. Conclusion

This study aimed to improve existing NTCP models to yield better prediction performance for late taste loss at 6 months post radiation therapy to improve overall quality of life. The inclusion of the tongue mucosa structure in NTCP models as a replacement of the oral cavity did not yield better performance than the reference model and a closer look must be taken at the accuracy of the derived structures. The neural network trained in this study has still a lot of potential for improvement. Many points that need addressing were discussed and should be included in future work. Overall, taste loss in HNC patients after radiotherapy treatment is a complex toxicity that must receive more research attention to be modelled and predicted more accurately.

Ethics Paragraph

Taste alteration is a commonly experienced side effect of radiotherapy treatment for HNC and has a huge impact on patients' lives by causing malnutrition, tube feeding and a drastic decrease in quality of life. A successful implementation of a reliable normal tissue complication probability (NTCP) model for the prediction of HNC radiotherapy related side effects like taste loss will cause an increase in quality of life for the growing number of HNC survivors. For this to be achieved currently available NTCP models must be improved upon to account for the large variability in patients treated.

Since one of the proposed solutions uses artificial intelligence (AI) in the form of a trained neural network to predict the patient's toxicity outcome, the implications of implementing AI in the medical field must be addressed. The European 'Ethics Guidelines for Trustworthy AI' (*37*) lists a set of key requirements needed to be met to classify as trustworthy AI. According to the guidelines, 'Human agency and oversight' needs to be maintained and for this project specifically, a human oversight over changes in patient characteristics and data acquisition that might affect the prediction outcome, as well as clinicians' agency to report non-conformance to predicted results must be ensured. Reliability and reproducibility fall under the requirement of 'Technical robustness and safety'. For the application in our research this means that firstly, the model created by AI must be tested and externally validated on a wide set of patient data, but also secondly, that the use conditions are clearly defined and communicated. What kind of patient characteristic, imaging manufacturer or imaging setting are allowed as inputs to the neural network needs to be very clear to the user. This is one of the biggest challenges in a field like the medical one, where research driven innovation changes clinical protocols and the technology used constantly. Publication by Morley et al. (*38*) and Naik et al. (*39*) emphasis the complexity of the wide variety of ethical issues that come with the introduction of AI into the medical field. The coming years will bring many new rules and regulations as well as general ethical assessment for medically used AI.

In conclusion, a prediction model developed by AI can be safe to use and have a positive impact on the quality of life of many HNC survivors. However, proper validation and human supervision during the implementation have to be ensured to allow for a reliable model.

References

1. J. Ferlay *et al.,* Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International journal of cancer*. **144**, 1941–1953 (2019), doi:10.1002/ijc.31937.
2. M. D. Mody, J. W. Rocco, S. S. Yom, R. I. Haddad, N. F. Saba, Head and neck cancer. *The Lancet*. **398**, 2289–2299 (2021), doi:10.1016/S0140-6736(21)01550-6.
3. *Hoofd-halskanker* (13/07/2022) (available at https://iknl.nl/kankersoorten/hoofd-halskanker).
4. *NKR Cijfers* (13/07/2022) (available at https://iknl.nl/nkr-cijfers).
5. B. Lacas *et al.,* Role of radiotherapy fractionation in head and neck cancers (MARCH): an updated meta-analysis. *The Lancet Oncology*. **18**, 1221–1237 (2017), doi:10.1016/S1470-2045(17)30458-8.
6. D. Pulte, H. Brenner, Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. *The Oncologist*. **15**, 994–1001 (2010), doi:10.1634/theoncologist.2009-0289.
7. B. M. Beadle *et al.,* Improved survival using intensity-modulated radiation therapy in head and neck cancers: a SEER-Medicare analysis. *Cancer*. **120**, 702–710 (2014), doi:10.1002/cncr.28372.
8. E. Benson, R. Li, D. Eisele, C. Fakhry, The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas. *Oral oncology*. **50**, 565–574 (2014), doi:10.1016/j.oraloncology.2013.09.008.
9. S. Marur, G. D'Souza, W. H. Westra, A. A. Forastiere, HPV-associated head and neck cancer: a virus-related cancer epidemic. *The Lancet Oncology*. **11**, 781–789 (2010), doi:10.1016/S1470-2045(10)70017-6.
10. B. J. M. Braakhuis, O. Visser, C. R. Leemans, Oral and oropharyngeal cancer in The Netherlands between 1989 and 2006: Increasing incidence, but not in young adults. *Oral oncology*. **45**, e85-9 (2009), doi:10.1016/j.oraloncology.2009.03.010.
11. J. A. Langendijk *et al.,* Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. **26**, 3770–3776 (2008), doi:10.1200/JCO.2007.14.6647.
12. B. A. Murphy, J. Gilbert, S. H. Ridner, Systemic and global toxicities of head and neck treatment. *Expert Review of Anticancer Therapy*. **7**, 1043–1053 (2007), doi:10.1586/14737140.7.7.1043.
13. T. S. Deshpande *et al.,* Radiation-Related Alterations of Taste Function in Patients With Head and Neck Cancer: a Systematic Review. *Current treatment options in oncology*. **19**, 72 (2018), doi:10.1007/s11864-018-0580-7.
14. H. Yamashita *et al.,* Taste dysfunction in patients receiving radiotherapy. *Head & neck*. **28**, 508–516 (2006), doi:10.1002/hed.20347.
15. T. T. Sio *et al.,* Intensity Modulated Proton Therapy Versus Intensity Modulated Photon Radiation Therapy for Oropharyngeal Cancer: First Comparative Results of Patient-Reported Outcomes. *International journal of radiation oncology, biology, physics*. **95**, 1107–1114 (2016), doi:10.1016/j.ijrobp.2016.02.044.
16. M. O'Neill *et al.,* Posttreatment quality-of-life assessment in patients with head and neck cancer treated with intensity-modulated radiation therapy. *American journal of clinical oncology*. **34**, 478–482 (2011), doi:10.1097/COC.0b013e3181f4759c.
17. L. van den Bosch *et al.,* Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: A new concept for individually optimised treatment. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. **157**, 147–154 (2021), doi:10.1016/j.radonc.2021.01.024.

18. S. Stieb *et al.,* Development and validation of a contouring guideline for the taste bud bearing tongue mucosa. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. **157**, 63–69 (2021), doi:10.1016/j.radonc.2020.11.012.
19. E. Vanetti *et al.,* Volumetric modulated arc radiotherapy for carcinomas of the oro-pharynx, hypo-pharynx and larynx: a treatment planning comparison with fixed field IMRT. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. **92**, 111–117 (2009), doi:10.1016/j.radonc.2008.12.008.
20. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature*. **521**, 436–444 (2015), doi:10.1038/nature14539.
21. J. Egger *et al.,* Medical deep learning-A systematic meta-review. *Computer methods and programs in biomedicine*. **221**, 106874 (2022), doi:10.1016/j.cmpb.2022.106874.
22. L. Boldrini, J.-E. Bibault, C. Masciocchi, Y. Shen, M.-I. Bittner, Deep Learning: A Review for the Radiation Oncologist. *Frontiers in Oncology*. **9**, 977 (2019), doi:10.3389/fonc.2019.00977.
23. A. L. Appelt, B. Elhaminia, A. Gooya, A. Gilbert, M. Nix, Deep Learning for Radiotherapy Outcome Prediction Using Dose Data - A Review. *Clinical oncology (Royal College of Radiologists (Great Britain))*. **34**, e87-e96 (2022), doi:10.1016/j.clon.2021.12.002.
24. X. Zhen *et al.,* Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Physics in medicine and biology*. **62**, 8246–8263 (2017), doi:10.1088/1361-6560/aa8d09.
25. B. Ibragimov *et al.,* Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy. *Medical physics*. **47**, 3721–3731 (2020), doi:10.1002/mp.14235.
26. K. Men *et al.,* A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial. *International journal of radiation oncology, biology, physics*. **105**, 440–447 (2019), doi:10.1016/j.ijrobp.2019.06.009.
27. L. van den Bosch *et al.,* Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. **148**, 151–156 (2020), doi:10.1016/j.radonc.2020.04.012.
28. *Deep Convolutional Neural Networks* (13/06/2022) (available at https://www.run.ai/guides/deep-learning-for-computer-vision/deep-convolutional-neural-networks).
29. U. Udofia, "Basic Overview of Convolutional Neural Network (CNN)". *DataSeries*, 13 February 2018 (13-2-2018) (available at https://medium.com/dataseries/basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17).
30. K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition". (10/12/2015; http://arxiv.org/pdf/1512.03385v1).
31. C. L. Brouwer *et al.,* CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. **117**, 83–90 (2015), doi:10.1016/j.radonc.2015.07.041.
32. S. Ruder, "An overview of gradient descent optimization algorithms". (15/09/2016; http://arxiv.org/pdf/1609.04747v2).
33. A. Chattopadhyay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, 839–847 (2018), doi:10.1109/WACV.2018.00097.
34. R. Matsuo, Role of saliva in the maintenance of taste sensitivity. *Critical reviews in oral biology and medicine : an official publication of the American Association of Oral Biologists*. **11**, 216–229 (2000), doi:10.1177/10454411000110020501.

35. M. Asif, A. Moore, N. Yarom, A. Popovtzer, The effect of radiotherapy on taste sensation in head and neck cancer patients - a prospective study. *Radiation oncology (London, England)*. **15**, 144 (2020), doi:10.1186/s13014-020-01578-4.
36. L. A. Barlow, Progress and renewal in gustation: new insights into taste bud development. *Development (Cambridge, England)*. **142**, 3620–3629 (2015), doi:10.1242/dev.120394.
37. *Ethics Guidelines for Trustworthy AI - FUTURIUM - European Commission* (NaN) (available at https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html).
38. J. Morley *et al.,* The ethics of AI in health care: A mapping review. *Social science & medicine (1982)*. **260**, 113172 (2020), doi:10.1016/j.socscimed.2020.113172.
39. N. Naik *et al.,* Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Frontiers in surgery*. **9**, 862322 (2022), doi:10.3389/fsurg.2022.862322.

Appendix A: Additional Tables

Table A1: Additional performance measures for the reference, DCNN and rCNN models. The limit is the value above which a prediction probability will be interpreted as an event. The true negative (TN), false negative (FN), false positive (FP) and true positive (TP) values are given in absolute patient numbers as well as in percentage based on the entire test set.

	Limit	F1	Sensitivity	Specificity	TN	FN	FP	TP
Reference	0,1	0.47	0.9	0.24	24 (17%)	4 (3%)	78 (55%)	36 (25%)
	0,2	0.54	0.8	0.55	56 (39%)	8 (6%)	46 (32%)	32 (23%)
	0,5	0.31	0.2	0.96	98 (69%)	32 (23%)	4 (3%)	8 (6%)
DCNN	0,1	0.44	1	0	0 (0%)	0 (0%)	102 (72%)	40 (28%)
	0,2	0.47	1	0.11	11 (8%)	0 (0%)	91 (64%)	40 (28%)
	0,5	0.48	0.68	0.55	56 (39%)	13 (9%)	46 (32%)	27 (19%)
rCNN	0,1	0.50	0.7	0.58	59 (42%)	12 (8%)	43 (30%)	28 (20%)
	0,2	0.51	0.6	0.70	71 (50%)	16 (11%)	31 (22%)	24 (17%)
	0,5	0.34	0.25	0.91	93 (65%)	30 (21%)	9 (6%)	10 (7%)

Appendix B: Additional Figures

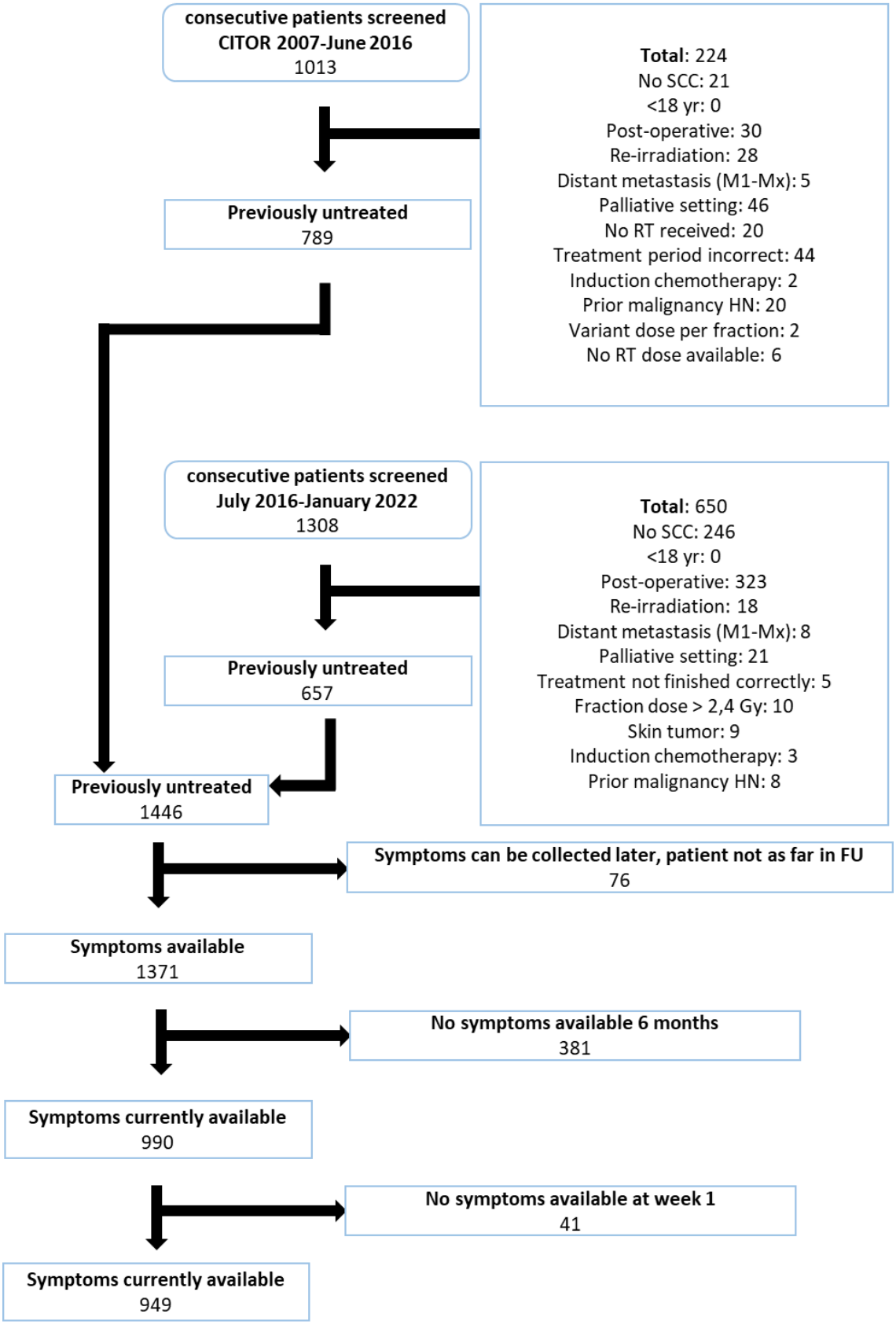

Figure A1: Patient inclusion diagram of patients suitable for this project. General exclusion criteria as well as missing endpoint data due to not completed follow up (FU) or missing symptom data are listed.

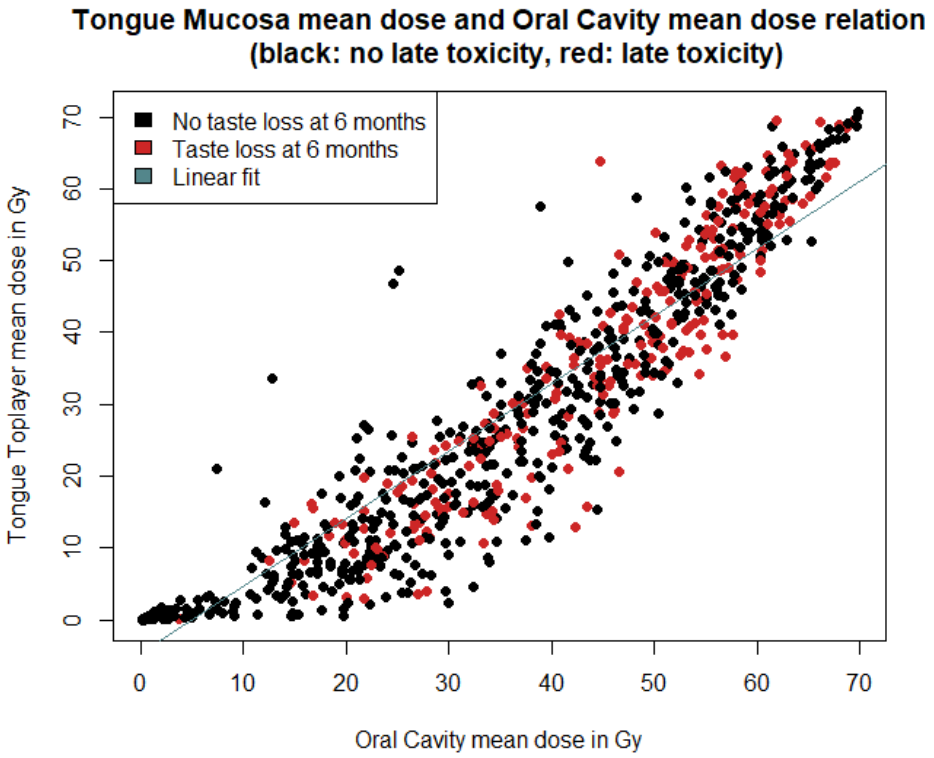

Figure A2: Scatter plot showing the relation between the tongue mucosa mean dose and the oral cavity mean dose, where patients experiencing taste loss at 6 months are shown in red. The linear fit with a slope of 0.94 and an intercept of -4.62 is shown in blue.

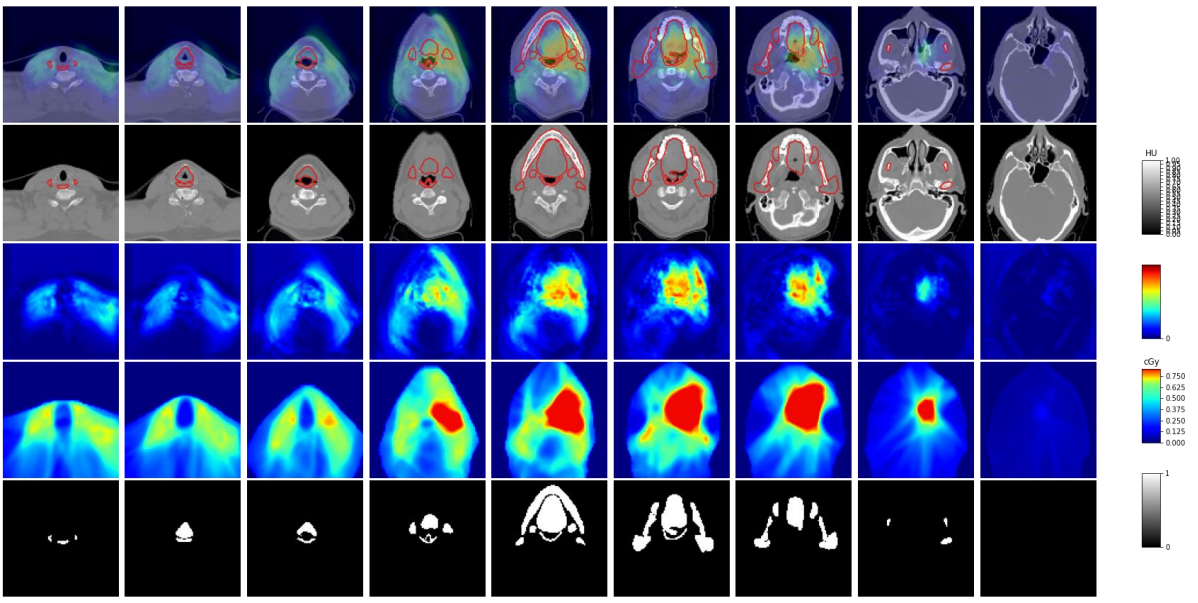

Figure A3: Attention map of DCNN showing high correlation between high dose regions and higher attention. From the top: Overlay of attention map on CT with RTSTRUCTs delineated in red; CT with RTSTRUCTs delineated in red; Attention map where red corresponds to higher correlation; Dose distribution where red corresponds to higher dose regions; binary RTSTRUCT mask with structures shown in white.

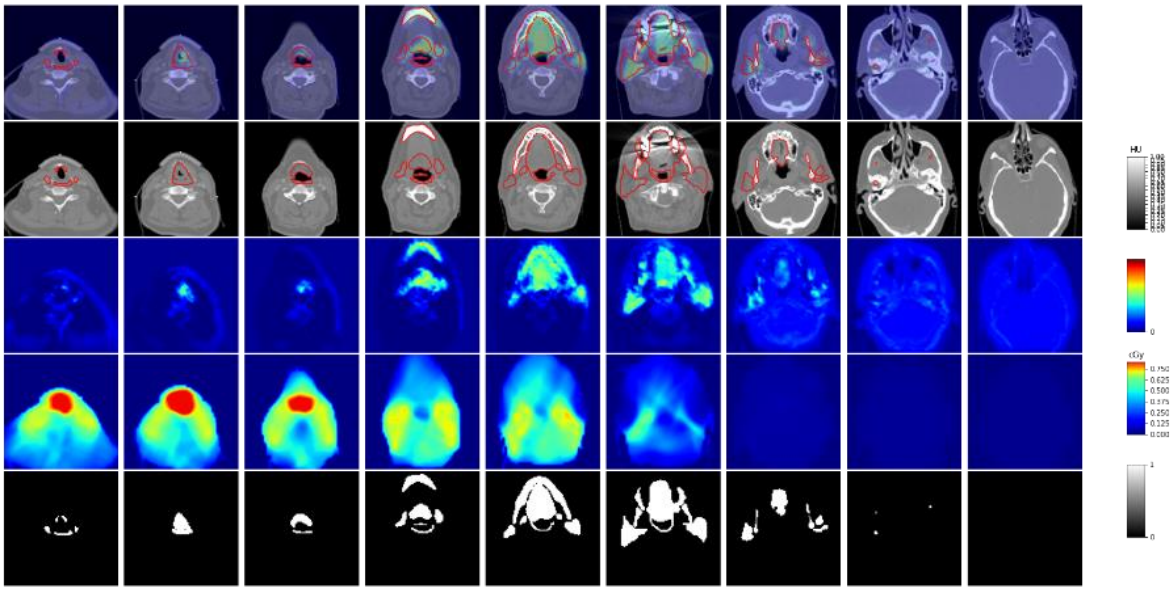

Figure A4: Attention map of rCNN showing high correlation between RTSTRUCT and higher attention. From the top: Overlay of attention map on CT with RTSTRUCTs delineated in red; CT with RTSTRUCTs delineated in red; Attention map where red corresponds to higher correlation; Dose distribution where red corresponds to higher dose regions; binary RTSTRUCT mask with structures shown in white.

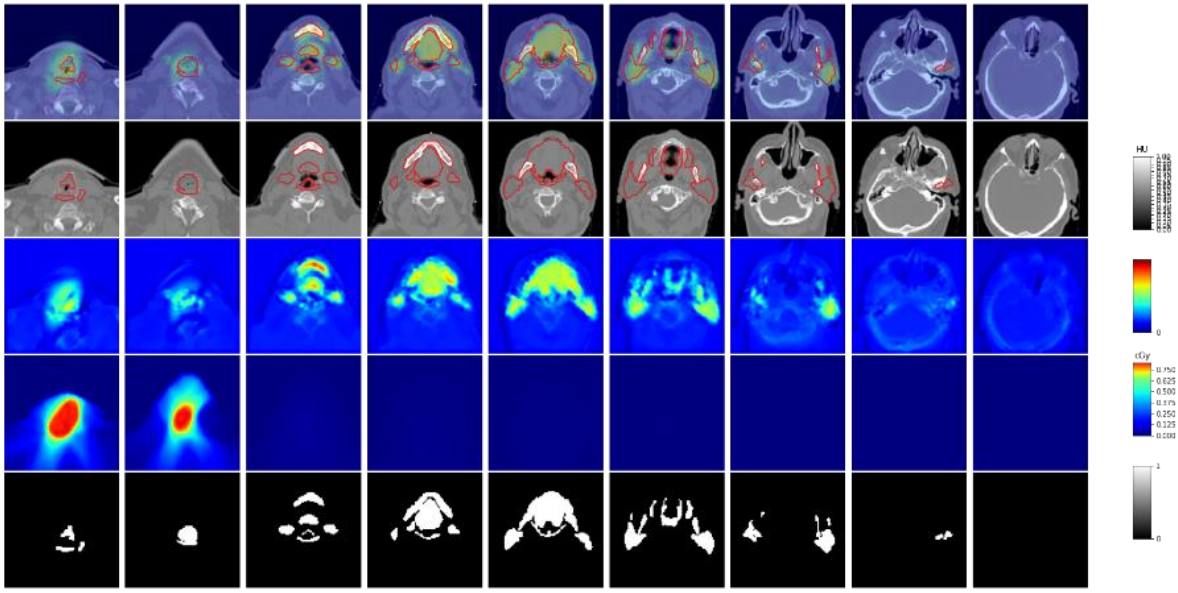

Figure A5: Attention map of rCNN showing high correlation between RTSTRUCT and higher attention as well as high dose region and higher attention. From the top: Overlay of attention map on CT with RTSTRUCTs delineated in red; CT with RTSTRUCTs delineated in red; Attention map where red corresponds to higher correlation; Dose distribution where red corresponds to higher dose regions; binary RTSTRUCT mask with structures shown in white.

Appendix C: Network Information

Table A2: Relevant starting network parameters before the systematic adjustment for both architectures.

	DCNN	rCNN
Input channels	3	3
Filters	[8, 8, 16, 16]	[8, 8, 16, 16]
Kernel sizes	[7, 5, 4, 3]	[7, 5, 4, 3]
Stride length	2	2
Optimizer function	Varied in first step	Varied in first step
Loss function	Cross entropy	Cross entropy
Label weights	[1, 3]	[1, 1]
Learning rate scheduler	Cosine	Cosine
T0	Varied in first step	Varied in first step
Maximum epochs	1000	1000
Batch size	8	8
Patience	35	35
Validation interval	1	1

Table A3: Final network parameters after the systematic adjustment of the best performing version of each architecture.

	DCNN	rCNN
Input channels	3	3
Filters	[8, 8, 16, 16]	[8, 8, 16, 16]
Kernel sizes	[7, 5, 4, 3]	[7, 5, 4, 3]
Stride length	2	2
Optimizer function	RMSProp	SGD
Loss function	Cross-entropy	F1
Label weights	[1, 3]	[1, 1]
Learning rate scheduler	Cosine	Cosine
Learning rate	3.76 e-06	7.56 e-04
T_0	16	40
Maximum epochs	1000	1000
Batch size	8	8
Early stopping (epochs)	35	35
Validation interval (epochs)	1	1

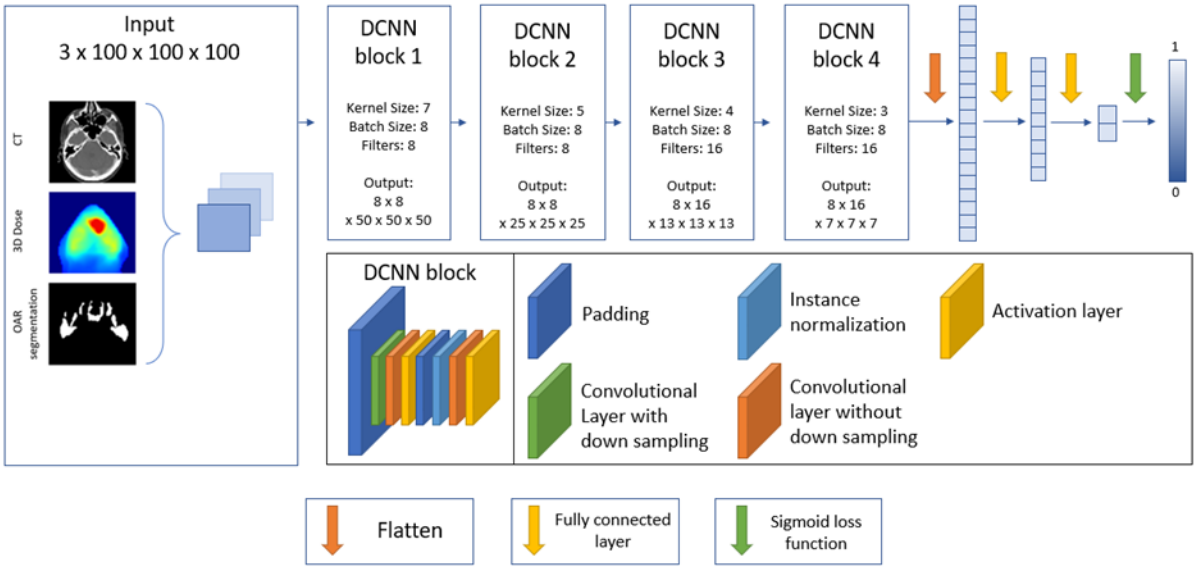

Figure A6: Schematic of the final DCNN network architecture.

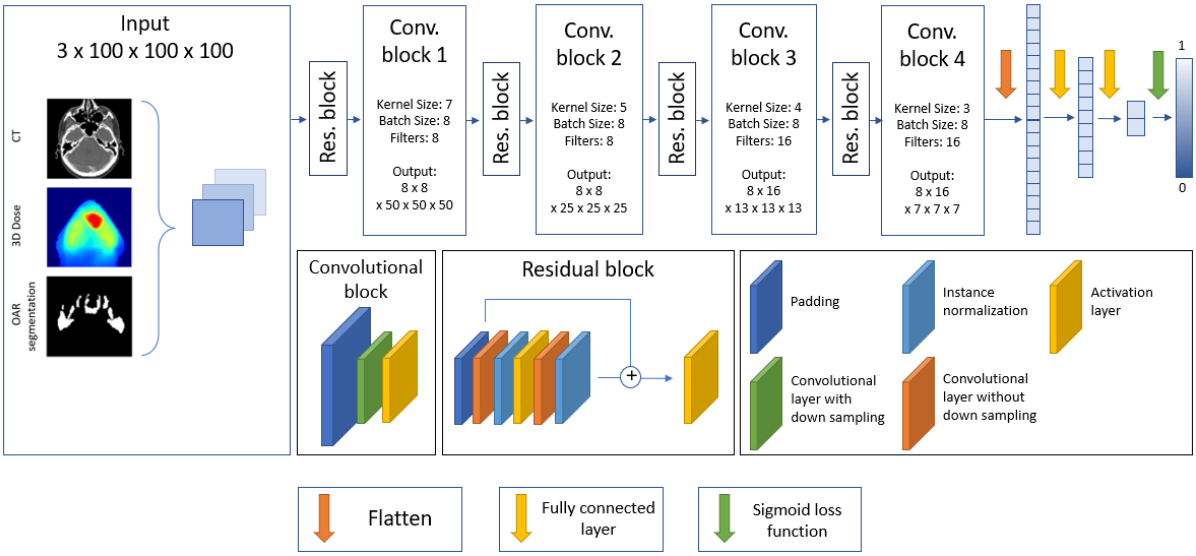

Figure A7: Schematic of the final rCNN network architecture.