Utrecht University

Graduate School of Life Sciences

- Postgraduate Epidemiology Program -

Master thesis

Master of Science

Development of NTCP-models to predict the presence of laryngeal edema six months after radiotherapy of head-and-neck cancer

Author:
PD Dr.med. Christina Heilmaier, MD, MBA, CAS
Lauersfortstrasse 12c
47804 Krefeld, Germany
SOLIS-ID 6331335

Examiner:
dr. ir. E. (Ewoud) Schuit
Daily supervisor:
dr. A.M. (Tuur) Leeuwenberg
Second reviewer:
dr. S.G. (Sjoerd) Elias

Table of contents

List of abbreviations

AIC	Akaike Information Criterion
CI	Confidence interval
CITOR	Comprehensive Individual Toxicity Risk
c-statistic	Concordance statistic
CRT	Conformal radiotherapy
CT	Computed tomography
CTCAE	Common Terminology Criteria for Adverse Events
CTx	Chemotherapy
DMAX	Maximum dose
DMEAN	Mean dose
DMIN	Minimum dose
DNA	Deoxyribonucleic acid
DVH	Dose-volume histograms
2D	Two-dimensional
3D	Three-dimensional
EORTC	European Organization for Research and Treatment of Cancer
EPP	Events-per-parameter
EUD	Equivalent uniform dose
Gy	Gray ($= \frac{1\,\mathrm{J}}{\mathrm{kg}}$)
HNC	Head-and-neck cancer
HPV	Human papilloma virus
ICD	International Classification of Diseases
IMRT	Intensity-modulated radiation therapy
LENT SOMA	Late Effects Normal Tissue Task Force-Subjective, Objective, Management, and Analytic
LEUD	Lyman equivalent uniform dose model
LOGEUD	Logit-equivalent uniform dose model
MeSH	Medical subject heading
MICE	Multivariate Imputation by Chained Equations

MRT	Magnetic resonance imaging
NTCP	Normal tissue complication probability
OAR	Organ-at-risk
OR	Odds ratio
PET-CT	Positron emission tomography-CT
PROGRESS	Prognosis Research Strategy
RCT	Randomized-clinical trial
RCTx	Radiochemotherapy
ROC	Receiver operating characteristic
RT	Radiotherapy
SCC	Squamous cell carcinoma
TiS	Carcinoma in situ
TNM system	TNM classification system of malignant tumors
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UICC	Union for International Cancer Control's
UMCG	University Medical Center Groningen
VIF	Variance-inflation-factor
VMAT	Volumetric modulated arc therapy

List of tables

List of figures

Abstract

Background: Photon-based radiation therapy is important in the treatment of head-and-neck cancer (HNC). However, potential post-radiogenic complications may have a long-lasting negative impact on the patient's quality of life. The current study focused on the presence of laryngeal edema 6 months after irradiation of HNC and had three main objectives: (I) external validation of the normal tissue complication probability (NTCP)-model published by Rancati based on the equivalent uniform dose (EUD), (II) development of new NTCP-models using a large training cohort, (III) external validation of the new models with an independent validation data set.

Methods: The training cohort consisted of n=750 and the validation cohort of n=395 patients. For both cohorts, a sub-cohort including patients with non-laryngeal HNC was formed (n=416 and n=227 participants, respectively). Multiple imputation was executed to compensate for missing values. Rancati's model performance was assessed with both the training and the validation cohort, determining the discrimination (area-under-the-curve, AUC) and calibration parameters (calibration intercept and slope). Following, candidate predictors for new models were selected, their correlation and multicollinearity was checked and the required sample size was calculated. New models were set up with stepwise regression ($p<0.20$), considering different (sub-)cohorts and candidate predictors to comprehensively answer the research question. The performance of the new models was analyzed and cross-validation (cv) was executed to account for optimism. Finally, external validation of the new models was performed.

Results: External validation of Rancati's model revealed a poor calibration and moderate model performance with the training data set (AUC=0.66; 95%-confidence interval (CI): 0.60-0.71). Analysis of 11 different NTCP-models showed that the model consisting of dose values (= mean dose to the glottic and supraglottic area = GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN), the treatment regimen (SEQUENCE), and clinical parameters (SMOKING and AGE) yielded the best result in estimating laryngeal edema in non-laryngeal HNC patients (cvAUC: 0.86 (0.74-0.97), calibration intercept: 0.06 (0.04-0.08), slope: 1.01 (1.0-1.01)). In contrast, only the dose values (GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN) turned out to be relevant in the sub-cohort of

laryngeal HNC patients (cvAUC: 0.66 (0.58-0.74), intercept: 0.24 (-0.19-0.46), slope: 1.17 (1.02-1.27)) and the total cohort (cvAUC: 0.75 (0.74-0.77), intercept: 0.04 (0.03-0.05), slope: 1.02 (1.02-1.02)), which could be attributed to differences in the clinical characteristics of the respective patients. Regardless of the patient cohort, the combination of GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN was superior to GLOTTIC DMAX and SUPRAGLOTTIC DMAX (= maximum dose to the glottic and supraglottic area) and the laryngeal EUD. Moderate to high correlation between the dose values was evident (range of correlation coefficients=0.69-0.92) but without significant multicollinearity as indicated by variance-inflation-factors of 1.01-2.16. Comparison of the external validation of Rancati's and the new NTCP-models with the validation data set revealed better performance parameters of the new models (e.g., AUC: 0.69-0.97 versus 0.51 for Rancati).

Conclusions: New NTCP-models relying on dosimetric and non-dosimetric variables achieve a better prediction of the risk of post-radiogenic laryngeal edema in HNC patients compared to recent models based entirely on dose data. This is an essential step towards individualized patient treatment.

Word count: 499

Laymen's summary

Head-and-neck cancers (HNCs) are relatively common and are usually treated with radiation therapy. However, as a result of the treatment, side effects can arise that affect the patient in his daily life. These complications include edema of the larynx, i.e., an increased accumulation of fluid, which, for example, impedes speech. For some time now, statistical models have been able to estimate how high the risk is of having a complication after radiation. In the first step of the present work, existing models for the prediction of laryngeal edema after irradiation of HNC were searched for. Two models were found, and one model was examined in detail with the participants of the current study as to how well the model can predict laryngeal edema. The model's performance was mediocre. In a second step, new models to predict the presence of laryngeal edema 6 months after the end of radiation were developed. Various characteristics of the 750 study patients and their tumors were considered as predictive factors. Several sub-cohorts were formed, e.g., a sub-cohort that included only participants with an HNC that did not directly affect the larynx (416 patients). A total of 11 different models were set up and examined in detail. While the pre-existing model only included predictive factors related to dose exposure levels, the new models also involved factors not directly related to dose. The best model was the one in which, in addition to the average dose with which the larynx was irradiated, the age of the patient, the smoking behavior, and any additional treatments (e.g., chemotherapy) carried out were listed, too. This model, which was developed in the sub-group of patients with a tumor outside the larynx, hence included non-dose data that are specific to a patient in addition to dose data. In contrast, in the sub-group with the laryngeal tumor patients and also in the total cohort, the model that contained only dose data was the best. It was thus shown that the optimal prediction model also depends on the subgroup in which it is applied, which is probably due to clinical characteristics of the respective patients. All new models were tested for their ability to predict the presence of laryngeal edema. Finally, the third step was to examine how good the new models are if they are used with unknown patients. This is because the models tend to work better with the patients they were developed with. For that purpose, the new models were assessed with other patients who were treated later

and partly at other hospitals. The same patients were also used to assess the performance of the currently existing models and the comparison with the new models showed that they allow for better prediction of laryngeal edema in HNC patients than the currently existing models. The results of the study are crucial and can help that the treatment of HNCs is further improved and increasingly tailored to the individual patient.

Word count: 487

1.) Introduction

1.1) Motivation of the study

Head-and-neck cancers (HNCs) account for about 2.8% of all cancers and are responsible for 1.8% of cancer deaths (1). A central part of the treatment of HNC is radiation therapy (RT), which can be carried out either alone or in combination with other therapies. Depending on the location of the primary tumor and the size of the radiation field, healthy tissue may also be damaged, which can lead to lasting functional impairments. Since this might negatively affect the patient's quality of life, radiation plans are drawn up before the start of therapy, in which the dose, the various radiation areas receive, can be read off. To further optimize and individually adapt patient's therapy, normal tissue complication probability (NTCP)-models have been developed and continuously improved. NTCP-models may be viewed as special type of prediction models that are used in RT. They can take on multiple tasks: (I) NTCP-models assist medical doctors to estimate a patient's prognosis for an outcome based on prediction factors, (II) they enable the identification of risk groups, (III) they guide clinical decision management (e.g., with regard to the choice of therapy or patient instructions), and (IV) they strengthen the knowledge about the development of a disease (2). Of notice, in the context of RT, NTCP-models predict the risk of post-therapeutic diseases but without drawing causal conclusions between the applied dose and the occurrence and strength of a potential complication (3,4).

During RT planning, NTCP-models allow for estimating the risk with which an organ-at-risk (OAR) located close to the radiation target might be damaged. This information is essential for the determination of the radiation dose that should be given. For some time now, NTCP-models have also been used for so-called counterfactual questions. They include, for example, considerations whether a patient could benefit from treatment with protons instead of photons (5,6). Before the introduction of NTCP-models, the decision for either photons or protons was primarily based on evidence from randomized controlled trials (RCTs). However, RCTs may be ethically problematic in some cases, are costly, and time-consuming and therefore may not be the optimal way to identify patients who are likely to benefit from another therapeutic option. Because of that, NTCP-models are increasingly being used as an alternative to answer

this question. The acceptance of NTCP-models is also reflected by the fact that in some European countries health insurance companies have accepted the predictions of the NTCP-models as basis for reimbursing the costs of proton therapy (7).

1.2) Aim of the study

The occurrence of laryngeal edema is a relatively common complication after RT of HNCs and was chosen as outcome of interest based on literature research and clinical aspects. The approach to determine the outcome of interest and to define the research question is explained in detail in the appendix (A1).

Given that the larynx fulfills multiple tasks in speaking, securing oxygenation, and nutrition, laryngeal edema can have devastating effects on the patient's quality of life and participation in social activities. The main objective of the present study was to develop NTCP-models to predict the presence of laryngeal edema 6 months after the end of RT for the treatment of HNCs. To comprehensively answer the research question, the following 3 steps were conducted:

- Step I: The NTCP-model by Rancati et al. (8), which is based on the equivalent uniform dose (EUD), served as starting point of the current study. As first step, Rancati's logit model applying the EUD was externally validated with both the training and the independently acquired validation data set of the current study.
- Step II: While Rancati's model relied solely on dose data, extensive non-dosimetric data of the participants (e.g., age, smoking behavior, other therapy options) was additionally available in the current study. Considering this add-on data, new NTCP-models were developed with the large training cohort in a second step. At this, several models were set up in different (sub-) cohorts taking into account different predictors to comprehensively answer the research question.
- Step III: External validation of the NTCP-models developed in step II was performed, using an independent validation cohort. These results were eventually compared with the findings of the external validation of Rancati's model executed with the same data set.

1.3) Background information

1.3.1) Statistics of head-and-neck cancers

The term HNC encompasses etiologically and histologically very different malignant neoplasms of the lip, oral cavity, nasal and paranasal sinuses, nasopharynx, oropharynx, larynx, and hypopharynx as well as tumors of the salivary glands. In most cases, however, HNC refers to tumors of the oro-/hypopharynx, the oral cavity, and the larynx, which hold the International Classification of Diseases (ICD)-10 code C00-C14 and C30-31. These tumors have a comparable etiology, arise from the mucosa of the upper aerodigestive tract, and predominate in the head-and-neck area. According to the Global Cancer Statistics 2020 published by the International Agency for Research on Cancer, n=931,931 new cases of HNCs were diagnosed in 2020, namely:

- lip, oral cavity: n=377,713
- larynx: n=184,615
- nasopharynx: n=133,354
- oropharynx: n=98,412
- hypopharynx: n=84,254
- salivary glands: n=53,583.

In the same time, n=467,125 new deaths due to HNCs were recorded (9). The incidence and mortality ratio of men compared to women lies between 2 and 3 for almost all HNC. Women have a cumulative incidence risk of 0.26 to suffer from tumors of the lip and oral cavity until age 74 with a cumulative mortality risk of 0.12. Compared to this, men hold a cumulative incidence risk of 0.68 and a cumulative mortality risk of 0.32, respectively, indicating that men incorporate a 2.7-fold higher cumulative risk. These gender-specific differences are also reflected in the age-standardized incidence rate per 100,000 inhabitants, which amounts to 3.1 for women and 6.0 for men in Northern Europe (9). Statistics depict that the incidence and prevalence of tumors of the lip and oral cavity varies considerably between countries and continents, but men are always significantly more often affected than women.

1.3.2) Risk factors of head-and-neck cancers

One underlying reason for the gender disparity in the occurrence of HNCs is that men still smoke tobacco and consume alcohol more frequently and in larger quantities than women, even if women have caught up in this regard in recent years. Moreover, men are more commonly affected by an infection with the oral human papilloma virus (HPV), which is another established risk factor for HNCs (10,11). Studies have proven an association of the presence of HPV and/or its surrogate marker p16 (cyclin-dependent kinase inhibitor 2A/multiple tumor suppressor 1) with the occurrence of squamous cell carcinoma (SCC) of the head-and-neck, particularly oropharyngeal cancer (12). The exact mechanism of HPV-mediated tumorigenesis has not yet been elucidated but appears to be different from that of tobacco.

In general, the incidence of HNCs raises with age and most patients are between 50 and 70 years old at the time of diagnosis. Lately, however, there was a trend towards HNC detection in younger patients, which likely is attributable to the increase of HPV-infections. The significant incline in the presence of HPV-associated cancer has led to an overall growth of the incidence of oropharyngeal cancer. This holds also true for the Netherlands with a gain of one third between 1990 and 2020 (n=2,075 and n=3,070, respectively) (13). In fact, a decline would have been expected for oropharyngeal cancer, as the number of smokers has shrunken in Western countries over the past two decades.

In addition to smoking and HPV-infection, excessive alcohol consumption is an important risk factor for HNCs. Almost all patients (>85%) have a medical history of long-standing, excessive consumption of tobacco and/or alcohol and both noxious substances show an almost linear association with HNC development (14,15). The joint abuse of the two substances even causes a significant synergistic/multiplicative reinforcement on the occurrence of HNCs, as highlighted in a recent study (10). The negative effect was most pronounced in tobacco abuse (odds ratio (OR)=11.6, 95% confidence interval (CI)=6.7-20.1), and to a lesser degree in consumption of five or more glasses alcohol per day (OR=2.7, 95% CI =1.2-4.7) as well as in case of an oral HPV-infections (OR=2.4, 95% CI=1.1-5.0) (10).

1.3.3) Staging of HNC

HNCs are usually diagnosed by biopsy and over 85% of HNCs are histologically SCCs, characterized by a rapid tumor proliferation with very short tumor doubling times (16). The tumors are characterized by often large volumes and early spread to regional lymph nodes, while hematogenic metastasis to the lungs, liver, and bones are usually confined to the advanced tumor stages.

Staging of HNCs follows the TNM classification system of malignant tumors (TNM-system) and depends on the size and location of the primary tumor (T), the number and size of cervical lymph node metastases (N), and the presence of distant metastases (M).

The T-stage is subdivided as follows:

- TiS= Carcinoma in situ; superficial tumor disease that grows only in the top cell layer of the skin or mucosa,
- T1-4= grading according to increasing size and extent of the tumor, classification differs depending on the location of the primary tumor.

Involvement of the lymph nodes is assigned as follows:

- N0= no lymph node infiltration,
- N1= 1 lymph node with malignant cells, which measures ≤3 cm at its greatest extent,
- N2= 1 or more lymph nodes with tumor infiltration of 3–6 cm in greatest diameter; a distinction is made as to whether the affected lymph nodes are on the same or on the contralateral side in relation to the primary tumor,
- N3= pathologic lymph nodes >6 cm in greatest extent.

With respect to distant metastases, a differentiation in M0 (no distant metastases present) and M1 (existence of at least one distant metastasis) is executed.

The TNM-system provides the basis for the HNC-staging of the Union for International Cancer Control's (UICC), which is given in Table 1 (17).

Stage	T	N	M
0	TiS (Carcinoma in situ)		
I	T1	N0	M0
II	T2	N0	M0
III	T1-3	N1	M0
	T3	N0	M0
IVa	T1-3	N2	M0
	T4a	N0-2	M0
IVb	T1-4a	N3	M0
	T4b	N1-3	M0
IVc	T1-4	N0-3	M1

Table 1. HNC-staging based on the UICC-classification, which considers the TNM-system (17).

After biopsy confirmation of the tumor diagnosis, staging usually relies on imaging modalities such as computed tomography (CT), and magnetic resonance imaging (MRI) as well as a supplementary positron emission tomography (PET)-CT.

1.3.4) Therapeutic options for the treatment of HNC

Various therapy options and combinations are available for the treatment of HNCs, encompassing surgery, radiotherapy (RT), and chemotherapy (CTx). The choice of therapy is grounded on the UICC stages (17). Commonly, surgical treatment is shorter and better tolerated, and hence may be preferred in the most comorbid patients. Yet, on the contrary, an operation is often associated with poorer vocal outcome (18). For tumors of the early stages I and II, primary RT constitutes an alternative to surgery, given its potential to preserve the organ and its functions. The total irradiation dose is mostly divided into small daily doses of 1.8–2.5 Gray (Gy), known as fractionation. This takes advantage of the fact that tumor cells normally have a lower ability to repair deoxyribonucleic acid (DNA)-damages than healthy tissue. Fractionation allows for a better protection and an increase in the maximum tolerated total dose of non-tumor tissue, enabling total doses of up to 80 Gy.

In more advanced stages, radiochemotherapy (RCTx), the combination of primary RT and CTx, or primary surgery followed by RT or RCTx could be options but may be accompanied by more severe side effects. Chemotherapeutic agents such as 5-

fluorouracil and cisplatin inherit a growth-inhibiting effect and intensify the radiation sensitivity of tumor cells, potentially improving the effect of irradiation.

Progress has not only been made in terms of the temporal application of the RT, but also of the type of radiation used. For several years now, some centers are able to execute not only standard photon therapy but also proton radiation. This exploits the fact that protons have another energy distribution curve compared to photons, which might be advantageous in some tumors.

1.3.5) RT in HNC treatment: technical aspects

Due to the close proximity of many critical structures, RT planning in HNCs is a major challenge to avoid lasting side effects and structural damages (19,20). Nowadays, planning is mostly accomplished using imaging data from CTs, sometimes after image fusion with MRI or PET data. Grounded on the image data, a 3D-model of the radiation field is created, which forms the basis for the precise RT outlining. However, RT-planning becomes more problematic if the tumor partially or even completely encloses an OAR. Particularly in these cases, the technical developments of recent years have paid off. While previously only 3D-conformal RT (3D-CRT) was deployable, nowadays with the introduction of intensity-modulated RT (IMRT) and volumetric modulated arc therapy (VMAT) more targeted and therefore gentler procedures for the adjacent OAR are applicable.

IMRT facilitates the modulation of the incident dose by setting up many small, irregularly shaped fields from many different directions of irradiation. This creates a sharp dose gradient between the target tissue and the surrounding non-tumor tissue. Through this, the radiation dose can be adapted even to complex HNCs adjacent to OARs.

VMAT can be regarded as an advanced option of IMRT and carries out 2 simultaneous processes: it radiates and rotates around the patient at the same time, which enables an even better dose distribution in shorter time (21). Particularly patients in severe pain benefit from VMAT given that for them, it is often impossible to lie still during a longer-lasting radiation. Yet, compared to IMRT, VMAT inherits a considerable higher planning automation, limiting the amount of manual changes and adaptations, which might have a negative impact on the protection of adjacent OARs. However, recent

studies unveiled that despite these limited adjustment options, an adequate protection of nearby OARs is warranted with the deployment of VMAT in HNC treatment (22,23).

1.3.6) RT in HNC treatment: physical aspects

Photons trigger various processes in the irradiated tissue. On the one hand, direct hits of photons in the tumor tissue arise, inhibiting the cell growth of essential biomolecules. On the other hand, photons trigger scattering processes in the radiation field through energy transfer, which takes place via the ionization of water molecules. This creates highly toxic free radicals that interact with cell components and damage the genetic material of tumor cells. If the resulting damage exceeds the ability of the tumor cells to repair themselves, they can no longer multiply because their mitosis is suppressed. The damage may even be so great that the cells die (apoptosis). To produce this devastating effect, the radiation-based harm must occur in close spatial and temporal proximity. The dose-effect relationship always exhibits a sigmoid form with first a slow, then faster increase and finally saturation. Since healthy tissue mostly divides at a slower rate than tumor cells, their dose-effect curve lies predominantly in the higher dose range.

1.3.7) Anatomy of the larynx

The larynx is a cartilaginous structure in the middle of the throat that functions as a connector between the throat (pharynx) and the windpipe (trachea). It separates the trachea from the esophagus and takes on three important functions: (I) separation of air and food passages, (II) regulation of the swallowing act, and (III) voice formation (phonation). Due to the densely packed anatomical conditions, the larynx stays in close contact with the hypopharynx (dorsal), the thyroid gland (ventral), and the vascular-nerve-sheath (lateral; common carotid artery, internal jugular vein, and vagus nerve).

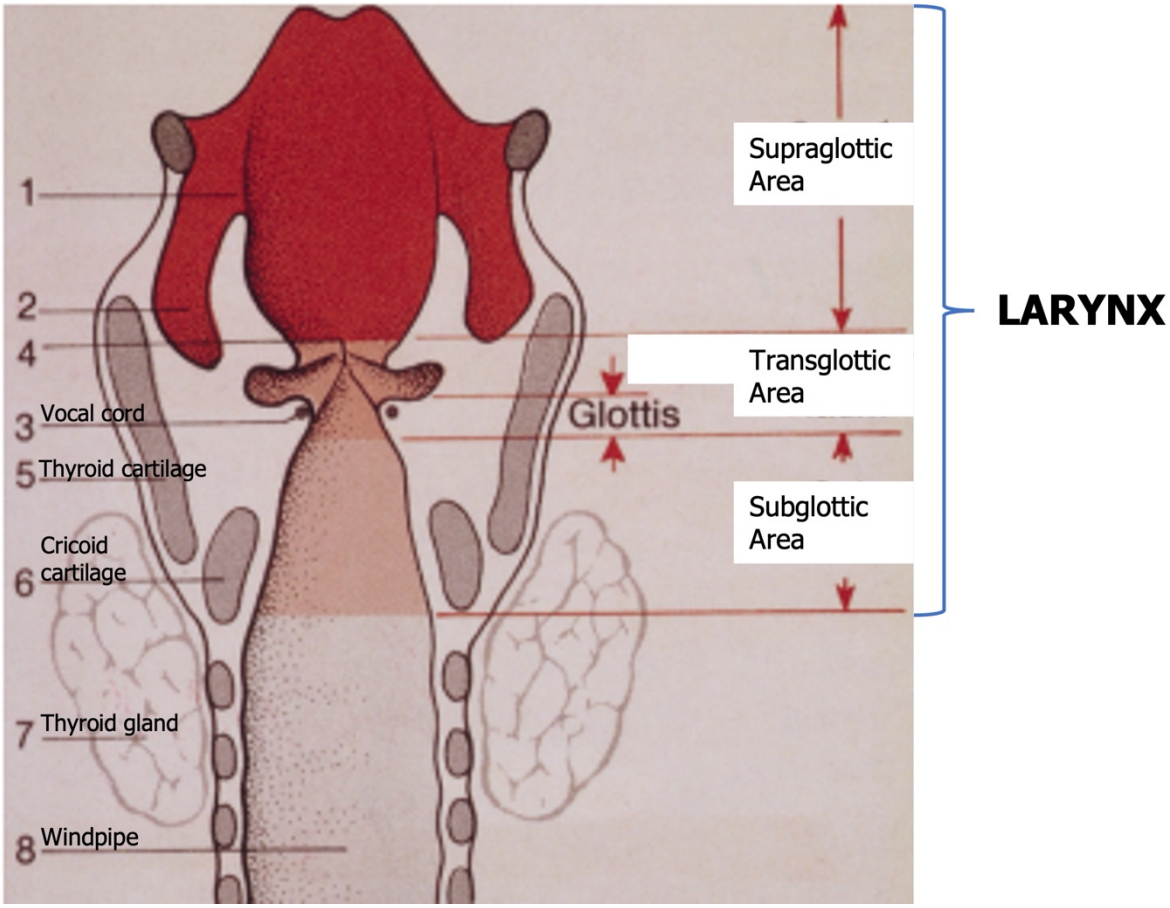

Figure 1. Schematic posterior view of the larynx. Adapted from (24).

The larynx is divided into an upper, middle, and lower third (figure 1): the upper level is the supraglottic area and extends from the entrance of the larynx to the pocket folds. The glottic area lies in the middle level of the larynx and contains the vocal cords, whose mobility is essential for sound formation. The space below the vocal cords up to the connection with the first tracheal clasp is the subglottic area and forms the lower end of the larynx.

1.3.8) Laryngeal edema

Increased accumulation of fluid in the lining of the larynx is denoted laryngeal edema. In the acute setting, it is oftentimes provoked by an external stimulus (e.g., inflammation, drug, foreign body) that induces a hypersecretion of the mucous membranes. Eventually, the excess amount of fluid produced is stored in the larynx tissue. In contrast, major structural changes are mostly evident in chronic laryngeal

edema. Damages to the walls of the small vessels of the mucous membrane result in the breakdown of the barrier that normally prevents the flow of fluids and solutes in the endothelial wall. As a consequence, the hydraulic conductivity and the osmotic reflection coefficient for plasma proteins as well as the lymphatic drainage system are disturbed (25). Thus, molecular changes in the vascular and the lymphatic systems play a role in the etiology of chronic edema.

The implications of laryngeal edema directly influence the patient's quality of life. The edema may lead to (chronic) hoarseness or aspiration of food that could be deposited in the lungs, ultimately leading to pneumonia. Moreover, a life-threatening edema with respiratory distress due to narrowing or blockage of the airways could appear. Finally, in severe cases, a tracheotomy may be required.

2.) Methods

2.1) Study design

The present study was conducted using data of a large prospective cohort study that took place in 3 medical centers in the Netherlands (Clinical trials NCT02435576). Among its objectives was the development of comprehensive individual toxicity risk (CITOR) profiles for HNC patients with definitive RT. Patients of the 3 medical centers were consecutively included in a data registration program as part of clinical routine. In addition to that, a prospective assessment of patient, tumor, and treatment characteristics as well as radiation-induced toxicity was performed in all participants.
Study results were reported as suggested by the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidance (26,27). As of today, the Dutch Medical Research Involving Human Subjects Act is not applicable to data collection in the context of routine clinical practice. For this reason, the responsible ethics committee waived the requirement for informed consent.

2.2) Study population

Patients who were treated at the University Medical Center Groningen (UMCG) between January 2007 and June 2016 formed the training cohort that was used (I) for the external validation of Rancati's model and (II) for the development of the new NTCP-models. A second group of patients formed the validation cohort and consisted of HNC patients from the Maastro Clinic (treated from May 2012 to June 2016), the Radiotherapy Institute Friesland (May 2014 to December 2016), and the UMCG (July 2016 to December 2017). The validation cohort was used for the external validation both of Rancati's model as well as of the new NTCP-models.
To account for missing values in the data sets, including missing data due to death (4), multiple imputation with the Multivariate Imputation by Chained Equations (MICE)-technique was carried out and repeated 10 times (28). Yet, imputation was only done with variables, in which at least 10 input data values existed as otherwise the risk of unreliable approximations would have been too high.

2.3) Eligibility criteria

To be eligible for the study, patients had to meet the following criteria: (I) biopsy-proven SCC of the head-and-neck (oral cavity, oropharynx, nasopharynx, hypopharynx, or larynx), (II) stage I-IV cancer without distant metastases (T_{1-4} N_{0-2} M_0), (III) primary therapy with RT, with or without concomitant CTx or cetuximab, (IV) no induction CTx, (V) no fraction dose higher than 2.4 Gy, (VI) no previous neck dissection, (VII) no previous therapeutic interventions for HNC (excluding laser resection of small glottic lesions), (VIII) no other malignancy in the last five years (excluding basal cell cancer or cervical carcinoma in situ), and (IX) no current co-existing tumors outside the head-and-neck region.

2.4) Patient treatment

Patients were treated in accordance with the Dutch guidelines for HNC (29). Until the end of 2007, most patients received 3D-CRT, which was progressively replaced by IMRT since 2008. Six-megavolt linear accelerators were applied to generate the radiation. Before the initiation of the RT, a contrast-enhanced planning CT scan was acquired in the supine position in all patients. In agreement with international consensus guidelines, 28 OARs were recontoured in all CT planning data sets (30). As additional preparatory measure before the start of the RT, all patients underwent a dental examination and, if necessary, a dental restoration was carried out. Furthermore, all participants received training on standardized oral care and mucositis prophylaxis.

The exact radiation protocol (e.g., fractionation plan, simultaneous CTx) depended on the primary tumor and the tumor stage. In general, younger patients (<70 years) with early-stage disease (stage I-II) received accelerated RT or fractionated RT. In patients with locally advanced disease (stage III-IV), platinum-based CTx (e.g., cisplatin) was administered concurrently with radiation to inhibit cell division. Contraindications for the administration of cisplatin comprised a status after apoplexy or myocardial infarction, intermittent claudication, neuropathy, loss of kidney function, or pre-existing severe hearing loss. In patients older than 70 years, conventional fractionation was adopted even in locally advanced stages, unless the patient's general state of

health permitted accelerated RT with weekly cetuximab. This therapy concept was also conducted in patients under the age of 70 when contraindications for CTx were depicted. Accelerated RT was given in 6 fractions per week, with a break of at least 6 hours between fractions.

In almost all patients, the regional cervical lymph nodes were also irradiated in addition to the primary tumour, even if a tumorous infiltration of the lymph nodes was not definitively proven. In these cases, the lymph nodes were treated in fractions of 1.55 Gy and received a total dose of 54.25 Gy. If the lymph nodes contained tumor cells, lymph node irradiation was carried out in fractions of 2 Gy together with the primary tumor up to a total dose of 70 Gy.

CTx was given concomitantly with conventional fractionated RT according to the following scheme: cisplatin 100 mg/m^2 on day 1, 22 and 43 or carboplatin on day 1 (300-350 mg/m^2 for 30 min intravenously) and 5-fluorouracil from day 1 to 4 as a continuous infusion (600 mg/m^2/24 h). In total, 3 cycles were undertaken at intervals of 3 weeks.

2.5) Measurement of predictors

Dose-volume-histograms (DVHs) are a fundamental part of RT planning given that they summarize 3D-dose distributions in a graphical 2D-format, facilitating the comparison of doses from different plans. In the context of DVHs, the term "volume" refers to all structures and targets of interest in the RT plan (e.g., primary tumor, adjacent OARs). To generate DVHs in the primary irradiation area and in the respective OAR, data of the contours and dose distributions in the current study were transferred to a suitable software program (VODCA Company: Viewer Version 4.2.2. and Database Version 4.1.1). Based on the software, calculation of the volume that received more than 5 Gy, 10 Gy up to a maximum of 75 Gy (V5, V10 up to a maximum of V75) was feasible. Furthermore, DVHs for the minimum dose (DMIN), maximum dose (DMAX), and mean dose (DMEAN) were computed and all DVH data was transferred to the database. This information was then combined with other patient-related variables, which may have an impact on the patient's prognosis.

After RT, participants were followed up for a maximum of 5 years and during this period, they were regularly called to the clinic for monitoring at specific times. During

these visits predefined questions related to the patient's quality of life, the health status, and the well-being were asked and relevant clinical examinations were executed.

2.6) Laryngeal edema

As already stated, post-radiogenic laryngeal edema was chosen as outcome of interest based on a comprehensive literature review and under consideration of clinical aspects (Appendix A1). Usually, the diagnosis of laryngeal edema was established by fiber-optic examinations and diagnostic findings were graded visually based on the Late Effects Normal Tissue Task Force-Subjective, Objective, Management, and Analytic (LENT SOMA)-table (table 2) (31).

Grade	**1**	**2**	**3**	**4**
Objective EDEMA	Arytenoids only	Arytenoids and aryepiglottic folds	Diffuse edema of supraglottis; airway adequate	Diffuse edema with significant narrowing of the airway (50% of normal)

Table 2. Overview of the LENT SOMA-table used for grading of laryngeal edema in fiberoptic examinations (31).

Thereafter, gradings from the LENT SOMA-table were transferred to the Common Terminology Criteria for Adverse Events (CTCAE) v5.0 (32):

- Grade 0: no edema
- Grade 1: asymptomatic edema
- Grade 2: symptomatic edema
- Grade 3: severe edema with respiratory distress
- Grade 4: life threatening edema.

Following the approach used in other studies (8,33–35), the presence of laryngeal edema of at least grade 2 was set as outcome of the present study during specification of the research question (Appendix A1).

2.7) NTCP-models by Rancati et al.

The literature research has revealed that up to now 2 NTCP-models for the prediction of post-radiogenic laryngeal edema have been published. Both models were developed by Rancati et al. and are based on the dose applied to the larynx.

- Model I: a Lyman model in which DVHs are reduced to the equivalent uniform dose (EUD) (LEUD model), and
- Model II: a logit model applying the EUD (LOGEUD model) (8).

The LOGEUD model was used almost exclusively in the studies on laryngeal edema published to date and therefore was chosen as basis for the current study (8,34,35). As first step of the external validation of Rancati's model with the training data set, the laryngeal EUD was calculated.

Due to the non-uniform dose distributions of DVHs, a straightforward comparison between RT plans is hampered. For that purpose, DVHs can be reduced to the EUD, which serves as an evaluation standard. The EUD corresponds to the absorbed dose which, when applied homogeneously to a tumor, has the same biological effect on the tumor as an inhomogeneous dose distribution (36). To determine the EUD, all existing dose data and the biological effects of fractionation are considered. Since this accounts for the overall effect on the tumor, the EUD can be viewed as a biologically weighted average of the dose (in voxels) across the tumor.

A more detailed explanation on the approximation of the EUD based on the DVHs as carried out in the present study is provided in the Appendix (A2).

Rancati et al. have performed the EUD transformation as follows (8):

$$EUD = \left(\sum_i v_i \cdot D_i^{\frac{1}{n}} \right)^n$$

D_i and v_i each represent a point of the differential DVH, where D_i stands for the dose from the DVH data and v_i indicates the fraction of the organ volume receiving a dose D_i. The sum is then performed over the entire DVH. The non-negative parameter n reflects the volumetric dependence of the dose-response relationship in every organ and influences the EUD as follows: the EUD tends toward the mean dose when n is close to 1 and dose non-uniformity is small. With large inhomogeneity, the EUD is in

the range of the minimum dose and with a very low n of ~0, the EUD reaches the maximum dose.

The EUD algorithm can be coupled with the logit formula to form the Logit-EUD (8):

$$NTCP\ (EUD) = \frac{1}{1 + \left(\frac{D_{50}}{D}\right)^{k}}$$

The dose-response relationship for healthy tissues is mirrored in the logit formula by D_{50} and k with D_{50} being the dose that comes along with 50% probability of damage if a certain fraction of the organ volume receives uniform irradiation. k can be calculated as $k = \frac{1.6}{\mathrm{m}}$ with m representing the steepest part of the slope of the response curve at D_{50}. The given values for the constants from the best-fit model from Rancati et al. (8) were taken over in EUD approximation in the current study (n=1.41, k=7.2, D_{50}=46.7).

After accomplishment of the EUD calculations, the performance of Rancati's LOGEUD model was assessed using both the 10 imputed data sets of the training cohort as well as the validation data set. For that purpose, the discrimination and calibration parameters were evaluated.

2.8) Performance measures

Several methods exist to evaluate the performance of a prediction model. In the current study, Nagelkerke's R^2, the discrimination, and the calibration parameters (and their corresponding CIs) were calculated as part of (I) the performance assessment of the newly developed NTCP-models and (II) during the external validation of Rancati's model and the new NTCP-models.

The general performance of a model is usually reflected by Nagelkerke's R^2, the so-called explained variation, which includes aspects of calibration and discrimination (38). More precisely, R^2 indicates the extent to which a predictor can explain the change in the response variable.

Discrimination is understood as how well the predictions can differentiate between patients with and without the outcome. In case of logistic regression models, the

concordance (c)-statistic is the most frequently used measure of performance (38). The discrimination can be represented graphically by means of the Receiver Operating Characteristic (ROC)-curve, in which the sensitivity (true positive rate) is plotted against 1-(false positive rate). From the plot the Area Under the ROC Curve (AUC) can be derived, which provides important information: the higher the AUC, the higher the model's discrimination ability (ideally 1.0). If a model has a binary outcome, the AUC and the c-statistic are identical (38).

Calibration refers to the correspondence between observed outcomes and predictions and can be illustrated graphically. In a calibration plot, the predictions are represented on the x-axis and the observed values are found on the y-axis, respectively. In case of binary outcomes, the y-axis ranges between 0 and 1. If the predictions are perfect, they should lie exactly on the 45°-line. Calibration performance can be quantitatively evaluated by the intercept and the slope of the calibration curve. The intercept states whether the predictions are systematically too low or too high (calibration-in-the-large) and should ideally be 0. The slope should ideally be 1, with values below 1 indicating an overfitting of the model (38). Overfitting refers to an over-adaptation of the model to the training data. When overfitting is present, model performance is poor if the model is confronted with other, so far unseen data, which is a common risk in iterative model development (38).

2.9) Sample size calculation

Determination of the sample size constitutes an important step in study planning and setting up a predictive model. It is crucial in reducing the risk of model overfitting and optimism, which may arise due to repeated testing in a small sample size (39). Riley et al. (40) have proposed approaches for binary, time-to-event, and continuous outcomes, respectively, that are available as pmsampsize package for R (41). The sample size calculation in the present study followed the four suggested criterions (B1-B4) and involved the parameters anticipated prevalence, number of predictors, and anticipated R^2. Furthermore, the shrinkage factor S was considered in the calculations, which should be ≥0.9 (40). To achieve this, targeting a shrinkage of ≤10% is recommended.

Even though publications have reported a prevalence of post-radiogenic laryngeal edema of up to 61% (8,42), a significantly lower prevalence of 5-10% was estimated in the present study, which was due to a different composition of the cohort as well as the more extensive eligibility criteria. The anticipated R^2 was expected as 0.1 or 0.15, respectively.

Table 3 provides a summary of the sample size calculations, when performed with the different variations of the parameters. Calculations were performed for n=12 candidate predictors, as this was the initial number of variables in the model development process. In the course of the development approach, the variables SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN were added; therefore, sample size calculations were repeated with n=14 candidate predictors.

			Criterion					
Anticipated		**Candidate predictors**	B1	B2	B3			B4
Prevalence	R^2	n	n	n	n	Events	EPP	n
0.05	0.1	14	73	269	1,189	60	4.2	799
0.1	0.1	14	139	374	1,189	119	8.5	547
0.05	0.15	14	73	269	768	39	2.7	776
0.1	0.15	14	139	374	768	77	5.5	531
0.05	0.1	12	73	233	1,020	51	4.3	685
0.1	0.1	12	139	324	1,020	102	8.5	469
0.05	0.15	12	73	233	659	33	2.7	666
0.1	0.15	12	139	324	659	66	5.5	456

Table 3. Overview of the sample size calculations following criterion B1-B4 as recommended by Riley et al. (40). Calculations were performed for n=12 and 14 candidate predictors, respectively, with an anticipated prevalence of 0.05 and 0.1, respectively, and an anticipated R^2 of 0.1 and 0.15, respectively.

EPP = Events-per-predictor

When calculating the sample size, the criterion that results in the highest necessary sample size is decisive. According to table 3, n=768 or n=1,189 participants were required if 14 candidate predictors were used in model development (with 12 predictors: n=659 and n=1,020, respectively). Due to the low prevalence of the outcome in the study population, the highest sample size could only be partially

achieved, but the lower criteria were met in all cases, so that overall, the study size can be assumed to be sufficient.

2.10) Correlation analysis and multicollinearity

The candidate predictors of the present study were selected based on literature research (Appendix A2 and A3) and availability in the data set.

In model development, it is important to consider potential implications of the interrelationships between the independent predictors, which is influenced by the size of the effect, the variance of the variable itself, the amount of data, the number of other variables in the model, and the size of the error variance. As a quantitative value, the correlation coefficient r shows the strength of the linear relationship between the variables.

In multicollinearity, two or more predictors are highly linearly associated with each other. The problems of multicollinearity are manifold (4). On the one hand, there is redundancy, i.e., two predictors could provide the same information about the outcome. This would ultimately lead to unreliable coefficients of the predictors and large standard errors of the coefficients, raising the risk of a type II error and reducing the prediction quality of the model (both in terms of reliability and accuracy). On the other hand, multicollinearity might cause that actually crucial predictors become unimportant because they have a collinear relationship to other predictors. This instability in predictor selection may lead to difficulties of applicability, especially for clinical prediction models. As a general rule, an absolute correlation coefficient of >0.8 between two or more predictors warrants a more detailed assessment of multicollinearity and eventually, variables might be excluded to correct for overfitting (4,43,61).

There are several ways to evaluate whether multicollinearity hampers model performance or not. It can first be checked whether there are significant changes in the estimated regression coefficients by adding or removing a predictor. Another indication of multicollinearity would be if a predictor behaves differently in univariable compared to multivariable regression, i.e., when the predictor has a coefficient significantly different from 0 in univariable regression but becomes insignificant in multivariable regression. Furthermore, the variance-inflation-factor (VIF) provides

important information, with a VIF of 5 or higher clearly pointing to multicollinearity. At this, the VIF can be interpreted as follows: a VIF of 5 indicates that the variance of the predictor's coefficient is 5 times higher than it would be without collinearity. It can easily be derived from Nagelkerke's R^2: VIF = $1/(1-R^2)$.

To account for potential correlation of the predictors and for multicollinearity, correlation coefficients and VIFs were computed for the predictors used in the new NTCP-models developed in the current study.

2.11) Development of new NTCP-models

At the beginning of the development process of the NTCP-models, a univariable logistic regression analysis of the candidate predictors was performed to examine the coefficients with their standard errors, the Akaike Information Criterion (AIC) as well as the concomitant statistical significance of the candidate predictors. This was followed by multivariable logistic regression procedures with the candidate variables, applying stepwise selection with rank order based on p-value. This approach has proven benefits for incorporating non-dosimetric factors into NTCP-models in the context of binary toxicity outcomes (44,45). The variables were selected using nested likelihood ratio test models, with a significance level of $p<0.2$. Thereafter, to adjust for overfitting and enhance calibration, especially if EPP is small, uniform shrinkage of the model regression coefficients was performed applying ridge regression (47).

Several NTCP-models encompassing changes in the candidate predictors, the outcome definition as well as the (sub-)cohort were developed to address research questions that arose in the course of the model evaluations. The rationale and background for the different models are explained in detail in the results section.

2.12) Internal validation of the new NTCP-models

According to the specifications of the TRIPOD-statement (46) and recommendations by Steyerberg et al. (38), the new NTCP-models were internally validated by 10-fold cross-validation. Cross-validation is a resampling method that uses different subsets of the data to train a model in different iterations (training set). The analysis is then validated on another subset (test set) and the process can be repeated several times

(commonly 10-times) to enhance stability of the cross-validation. The goal is to determine how good the model's ability is to predict new data that was not used in the estimation.

2.13) Evaluation performance measures of the new NTCP-models

The R^2 as well as the calibration and discrimination parameters were independently assessed in the 10 imputed data sets and results were averaged. During this process, it had to be considered that the data was not actually observed, but rather estimated as part of the imputation procedure, which had an inflating effect on precision. To correct for this, the regression coefficients and standard errors of the predictors were pooled using Rubin's rule (48).

2.14) External validation of the new NTCP-models

To get an idea on the generalizability of the new NTCP-models, their calibration and discrimination abilities were evaluated using the validation data set that was acquired independently of the training data set. Since Rancati's model was validated externally with both the training data set and the validation data set, a direct comparison of the performance of the new NTCP-models and Rancati's model was possible.

2.15) Data analyses

All analyses were performed using the software R, version 4.1.1. The following R-packages were applied:

boot, bravo, CalibrationCurves, car, caret, dplyr, Ecdat, GGally, ggplot2, gplots, glmnet, GmAMisc, kimisc, lattice, lmtest, Metrics, mice, pmsamplesize, pROC, randomForest, readxl, riskRegression, rmda, rms, ROCit, ROCR, stats, tidyverse.

3.) Results

3.1) Training data set

The training data set encompassed n=750 patients. The sex ratio of men versus women was 3, being in line with the known sex disparities in HNCs. Moreover, almost all patients (88%) had a history of or were still smoking. More advanced tumor stages (stage III or higher) were evident in more than two-third of the patients, and all of these participants had at least one proven lymph node metastasis. Being an exclusion criterion for study participation, no organ metastases were present in any patient. HPV was merely detected in 34 participants; however, HPV assessment was solely executed in a total of n=42 patients. Taking the evaluation rate into account, HPV was detectable in 34/42 (81%) of the patients.

As previously mentioned, it is virtually impossible to clinically distinguish whether a laryngeal edema occurred as an immediate complication of the laryngeal tumor or as sequela of irradiation. For this reason, a sub-cohort was formed, which only included patients with non-laryngeal HNCs and consisted of n=416 patients. The juxtaposition of the sub-cohort with the total cohort showed that more participants in the sub-cohort suffered from an advanced tumor disease (stage IVa and IVb; 76% in sub-cohort versus 51.2% in total cohort) and that the men-women-ratio was somewhat smaller (2:1 in the sub-cohort versus 3:1 in the total cohort). The other clinical parameters were proportionate in the sub-cohort and total cohort of the training data set.

Different treatment regimens were conducted depending on the tumor stage as well as the patient's age and general condition. As a consequence of relatively more participants with advanced tumor disease in the sub-cohort, differences in the treatment modalities became obvious. Significantly more patients received RTCx or the combined administration of RT with cetuximab, while RT alone was executed less frequently in the sub-cohort of non-laryngeal patients.

Factor	Level	Total training cohort		Sub-cohort	
		No. of patients	%	No. of patients	%
Sex	Man	560	74.7	279	67.1
	Woman	190	25.3	137	32.9
Age	Mean	63		61.5	
	Range	30-92		30-92	
Smoking	Yes (previously)	254	33.9	125	30.0
	Yes (still smoking)	402	53.6	232	55.8
	No	94	12.5	59	14.2
Tumor location	Oral cavity	44	5.9	44	10.6
	Oropharynx	271	36.1	271	65.1
	Nasopharynx	30	4.0	30	7.2
	Hypopharynx	71	9.5	71	17.1
	Larynx	334	44.5	0	0
N-stage	N0	333	44.4	75	18.0
	N1-3	417	55.6	341	82.0
Tumor stage	Stage I	85	11.3	12	2.9
	Stage II	148	19.7	37	8.9
	Stage III	133	17.7	51	12.3
	Stage IVa	322	42.9	259	62.3
	Stage IVb	62	8.3	57	13.7
Histology	SCC	709	94.5	386	92.8
	Nasopharyngeal carcinoma	30	4.0	30	7.2
	Tis	1	0.1	0	0
	Not specified	10	1.3	0	0
HPV	Negative	8	1.1	4	13.3
	Positive	34	4.5	26	86.7
	Not assessed	708	94.4	386	92.8
RT technique	3D-CRT	86	11.5	19	4.6
	IMRT	546	72.8	341	82.0
	VMAT	118	15.7	56	13.4
Treatment modality	CRT	149	19.9	42	10.1
	Accelerated RT	294	39.2	101	24.3
	RCTx	242	32.3	213	51.2
	Accelerated RT with cetuximab	65	8.7	60	14.4
Neck irradiation	No	147	19.6	45	10.8
	Unilateral	18	2.4	4	0.9
	Both sides	585	78	367	88.3

Table 4. Descriptive statistics of the total training cohort and the sub-cohort of non-laryngeal HNC participants: basic characteristics of the patients and the tumors as well as therapeutic options.

3.2) Validation data set

The validation cohort comprised n=395 patients. Comparison of the training and validation data set revealed no significant differences regarding age, sex, tumor location, tumor stage, histology, and neck irradiation. However, considerable discrepancies in the treatment were evident due to technical advancements. VMAT was almost exclusively applied in the validation cohort, while participants of the training data set mainly underwent IMRT. Regarding the treatment modality, in relation more CRT and fewer accelerated RT was performed in the validation cohort than in the training cohort. The juxtaposition of HPV detection (or its surrogate p16) in the two cohorts was hampered by a high number of missings in the training cohort, in which only in 6% an HPV test was executed. In contrast to that, HPV was examined in more than 86% of the patients in the validation cohort.

To ensure the same basic setting in the training and validation cohort, respectively, patients, in whom the primary tumor was in the larynx, were excluded from the validation cohort and a sub-cohort was built. This reduced the number of patients in the sub-cohort of the validation data set to n=227. Table 5 summarizes the characteristics of the patients and their therapeutic interventions in the total validation data set and the sub-cohort.

Compared to the non-laryngeal sub-cohort of the training data set, some differences became evident, e.g., regarding age with older patients in the validation sub-cohort ($p<0.05$) or nicotine abuse with more smokers among the patients in the training sub-cohort. Furthermore, overall patients in the validation group suffered from a less advanced tumor stage and almost all of them were treated with VMAT.

Factor	Level	Total validation cohort		Sub-cohort	
		No. of patients	%	No. of patients	%
Sex	Man	290	73.4	163	71.8
	Women	105	26.6	64	28.2
Age	Mean	63.3		63.5	
	Range	44-88		40-93	
Smoking	Yes (previously)	88	22.2	50	22.0
	Yes (still smoking)	82	20.8	94	41.4
	No	92	23.3	59	26.0
	Not assessed	133	33.7	24	10.6
Tumor location	Oral cavity	22	5.6	22	9.7
	Oropharynx	140	35.4	140	61.7
	Nasopharynx	15	3.8	15	6.6
	Hypopharynx	50	12.7	50	22.0
	Larynx	168	42.5	0	0
N-stage	N0	190	48.1	56	24.7
	N1-3	205	51.9	171	75.3
Tumor stage	Stage 0	1	0.3	1	0.4
	Stage I	55	13.9	30	13.2
	Stage II	58	14.7	17	7.5
	Stage III	90	22.8	60	26.4
	Stage IVa	177	44.8	111	48.9
	Stage IVb	14	3.5	8	3.5
Histology	SCC	383	97.0	219	96.5
	Follicular carcinoma	2	0.5	1	0.4
	Non-specified	1	0.3	2	0.9
	Missing data	9	2.3	5	2.2
HPV	Negative	285	72.2	78	34.3
	Positive	56	14.2	53	23.3
RT technique	3D-CRT	6	1.5	2	0.9
	IMRT	7	1.8	4	1.8
	VMAT	382	96.7	221	97.4
Treatment modality	CRT	126	31.9	63	27.8
	Accelerated RT	110	27.8	39	17.2
	RTCx	134	33.9	108	47.6
	Accelerated RT with cetuximab	25	6.3	17	7.5
Neck irradiation	No	66	16.7	1	0.4
	Unilateral	18	4.6	15	6.6
	Both sides	311	78.7	211	93.0

Table 5. Available data on smoking behavior, tumor location, histology, and treatment characteristics of the patients of the total validation cohort and the sub-cohort of non-laryngeal HNC participants.

3.3) Laryngeal edema

In the training sub-cohort, a clinically relevant laryngeal edema was present in 7 of the 416 (1.7%) patients at baseline and in 29 of the 416 (7.0%) participants 6 months after RT; thus, a significant increase was evident ($p<0.05$).

Data analysis in the validation data set was hampered by a high number of missing values particularly regarding laryngeal edema, which may have been due to several reasons (e.g., not assessed as the patients had no complaints or because the patient's general health condition did not allow an invasive endoscopic examination). At baseline, information on laryngeal edema was available from n=130/227 (57.3%) patients, of whom 4 (3.1%) had an edema grade 2 or higher, while 5 (3.8%) suffered from an asymptomatic edema, and 121 (80.1%) had no signs of disease. Six months later, n=152/227 (67.0%) patients have been examined, of whom 2 (1.3%) demonstrated a laryngeal edema grade 2 or more, while 26 (17.1%) had an asymptomatic edema, and in 124 (81.6%) participants the laryngeal mucous membrane showed no lesion. Hence, in the validation data set considerably more patients suffered from an asymptomatic edema 6 months after RT than at baseline, but fewer participants had a symptomatic post-radiogenic edema at the index date.

3.4) External validation of Rancati's LOGEUD-model

Due to disparities in data availability, the EUD calculation was executed in two different variants during the external validation of Rancati's LOGEUD model. In the first variant, laryngeal EUD was calculated, including all dose data of the training set (V05, V10, V15, V20, V25, V30, V35, V40, V45, V50, V55, V60, V65, V70, V75, V80, V85). This resulted in a moderate discrimination performance of Rancati's model, with an AUC of 0.66 (95% CI: 0.60-0.71) and a low R^2 of 0.05 (95% CI: 0.04-0.05).

Compared to the training cohort, dose data was recorded in a less elaborate manner in the validation data set (registered dose data: V05, V10, V20, V30, V40, V50, V60, V70). To ground EUD calculation on the same dose data in the training and validation data set, a second variant for laryngeal EUD approximation was set up, which relied solely on the dose data provided in the validation cohort. Applying the second variant for the evaluation of the discrimination performance of Rancati's model with the

training data set resulted in a small decrease of the AUC (0.64) with slightly wider CIs (0.56-0.73) and an unchanged R^2 of 0.05 (95% CI: 0.04-0.05).

Calibration assessment with the training data set showed a poor calibration of Rancati's LOGEUD model (regardless of whether the first or second variant of EUD approximation was considered). Quantification of the model's calibration revealed an intercept of -0.84 (95% CI: -10.73-9.05), an R^2 of 0.06 (95% CI: 0.05-0.06), and a slope of 1.21 (95% CI: -0.77-3.19).

Using the validation data set for the external validation of Rancati's model resulted in an AUC of 0.51 (95% CI: 0.39-0.62), an R^2 of 0.21 (95% CI: 0.2-0.22), a calibration intercept of -0.47 (95% CI: -0.52-(-0.3)), and a calibration slope of 0.99 (95% CI: 0.97-1.0).

3.5) Development of new NTCP-models

3.5.1) Selection of candidate predictors

The candidate predictors chosen are summarized in table 6. Bae et al. (49) examined possible risk factors using a univariate analysis and were able to identify tumor localization, T- and N-classifications, overall stage, pathological differentiation, and CTx as significant predictors, but only T-stage was found to be an independent predictive factor in a subsequent multivariable analysis. Because of that the variable T-stage (T-STAGE) was picked as candidate predictor in the present study.

In addition, tumor location (LOCATION), histology (HISTOLOGY), and CTx were selected as candidate predictors. For the variable CTx the temporal sequence between the application of CTx and RT was taken into account, too (e.g., induction CTx, concomitant RTCx) and this variable was denoted SEQUENCE. Furthermore, the RT technique (TECHNIQUE) was chosen as candidate predictor since improvements in the hardware also may have an impact on the presence of post-radiogenic diseases.

Machay et al. had detected a link between age and the occurrence of laryngeal edema in their study (50) and similar findings were seen after the RT of other entities, in which age always boosted the development of post-radiogenic side effects (= the higher the age, the higher the change of post-radiogenic complications) (51). Because of that and due to clinical considerations, AGE was added as candidate variable in

model development. Along with alcohol consumption, cigarette smoking has been identified as the most common risk factor for the development of HNC (52,53) and leads to permanent changes in the mucous membranes, especially in the mouth and throat area. Two studies have shown that smoking likewise promotes the development of post-radiogenic edema (54,55), which might be due to the wide-ranging alterations of immunological functions caused by nicotine and other substances contained in cigarettes, i.e., alternation of innate and adaptive immune responses (56). Furthermore, cigarettes have a negative impact on tissue oxygenation and induce hypoxia, thereby leading to the activation of several transcription factors such as vascular endothelial growth factor that supports the formation of tumor vessels. Moreover, hypoxia may alter the oxygen-dependent effects of RT and change drug delivery to tumor cells during CTx (57,58). As implication of a decreased response to RT, higher local doses need to be applied to destroy the tumor cells, which is linked with an increased risk for the adjacent OARs. Accordingly, SMOKING was added as candidate predictor, but for simplification, the two factor levels "current smoking" and "smoking in the past" were combined to one level (SMOKING: "Yes").

Assuming that a patient would have a growing risk of being diagnosed with laryngeal edema 6 months after RT, if already at baseline a laryngeal edema was present, the variable BASELINE EDEMA was included as a candidate predictor. Moreover, HPV was added as candidate variable as recently a causal relationship between HPV-infection and various types of HNC has been demonstrated (11,59,60). Furthermore, Jeans et al. verified a significant effect of HPV-infection not only on the occurrence of HNC but on the prediction of laryngeal edema, too (42).

In addition to the laryngeal EUD, which was adopted from Rancati's model, the maximum dose applied to the supraglottic (SUPRAGLOTTIC DMAX) and glottic area (GLOTTIC DMAX) were selected as candidate predictors. In the further course of the analysis, the 2 variables SUPRAGLOTTIC DMEAN and GLOTTIC MEAN, the mean dose to the supraglottic and glottic area, respectively, were added. At the beginning, these 2 factors were not chosen as it was guessed that they inherit roughly the same information as the EUD, but this assumption was later revised. Ultimately, a total of 14 parameters were involved in the model development process (table 6).

Abbreviation	Explanation	Factor levels
EUD	EUD of the glottic and supraglottic area together	numerical
BASELINE EDEMA	laryngeal edema at baseline	grade 0 (no edema); grade 1 (asymptomatic); grade 2 (symptomatic); grade 3 (severe); grade 4 (life threatening); not assessed
SMOKING	smoking status	no; yes; not known
HPV	presence of HPV	negative; positive; not assessed
SEQUENCE	sequence of CTx or cetuximab with RT	only RT; concomitant RCTx; accelerated RT with cetuximab, concomitant RTCx + adjuvant CTx; induction CTx; induction + concomitant CTx
LOCATION	localisation of primary tumor	oral cavity; oropharynx; nasopharynx; hypopharynx; larynx; paranasal sinus/ nasal cavity; salivary glands; ear; unknown primary; thyroid gland; skin; miscellaneous
T-STAGE	final T-stage of primary tumor	Tis; T0; Tis; T1; T2; T3; T4; T4a; T4b
AGE	age at time of diagnosis	numerical
HISTOLOGY		SCC; nasopharynx carcinoma; verrucous carcinoma; sarcoma; carcinoma-in-situ; medullary carcinoma; follicular carcinoma; other histology
SUPRAGLOTTIC DMAX	Maximum dose to the supraglottic area	numerical
GLOTTIC DMAX	Maximum dose to the glottic area	numerical
SUPRAGLOTTIC DMEAN	Mean dose to the supra-glottic area	numerical
GLOTTIC DMEAN	Mean dose to the glottic area	numerical
TECHNIQUE		3D-CRT; IMRT; VMAT; IMRT+VMAT

Table 6. Overview of the 14 candidate predictors and their levels involved in the model development process.

3.5.2) Correlation and multicollinearity of variables

A total of 6 numerical parameters were selected as candidate variables, namely EUD, AGE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX, SUPRAGLOTTIC DMEAN, and GLOTTIC DMEAN. To analyze their correlation, the correlation coefficients and correlation matrix were calculated using the data set of the sub-cohort containing the n=416 patients with non-laryngeal HNCs (figure 2).

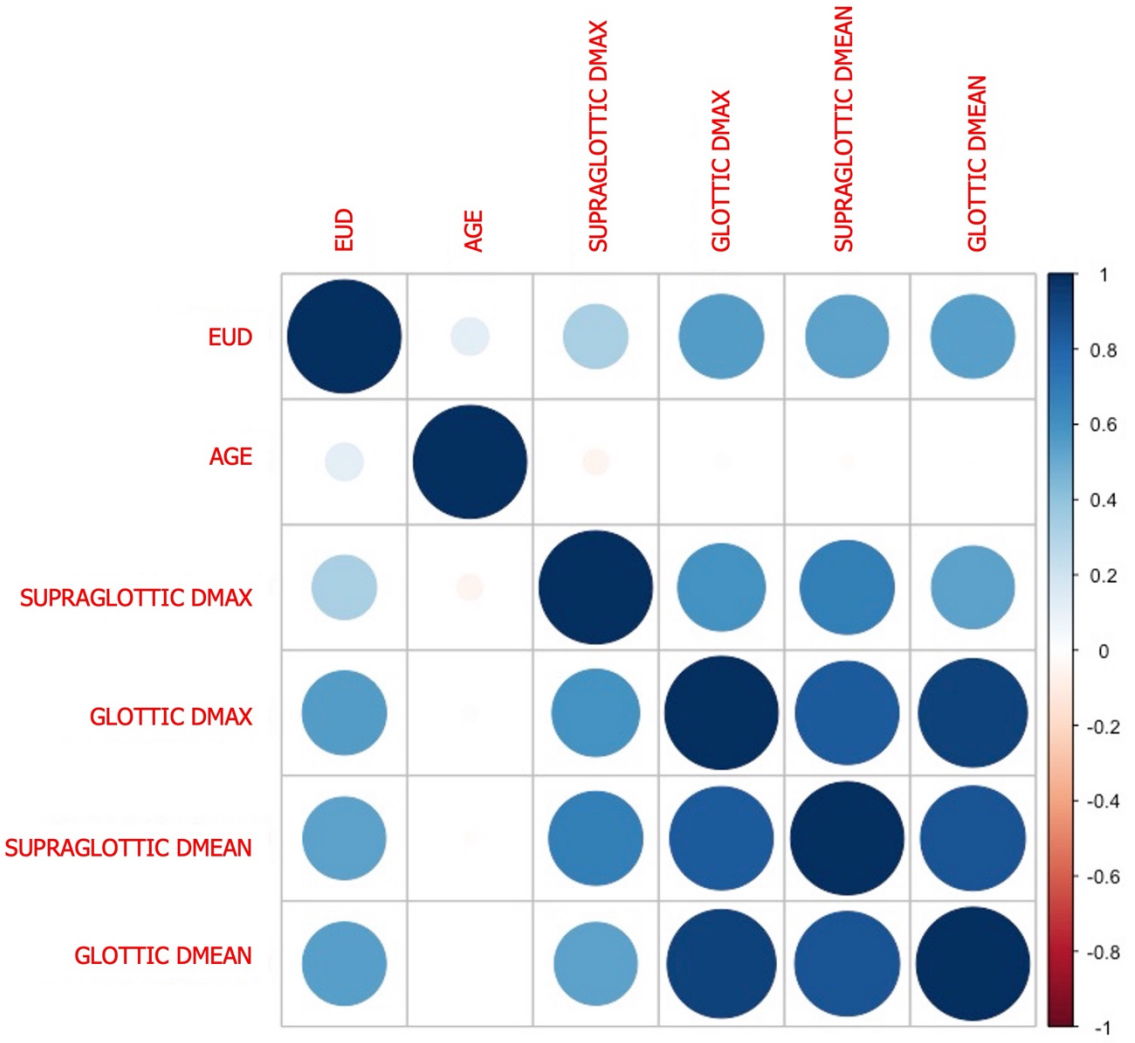

Figure 2. The correlation matrix illustrates the strong correlation of the dose-related parameters, particularly of GLOTTIC DMAX and GLOTTIC DMEAN.

As expected, the dose-related variables exhibited a high correlation, which was most pronounced between GLOTTIC DMEAN and GLOTTIC DMAX ($r=0.92$, $p<0.001$), followed by GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN ($r=0.86$, $p<0.001$). In

comparison, correlation of GLOTTIC DMAX and SUPRAGLOTTIC DMAX was significantly lower and only of moderate strength ($r=0.6$, $p<0.001$). Interestingly, the EUD correlated only weakly to moderately with the other dose-related values ($r=0.32$-0.55). AGE showed a feeble correlation with the EUD ($r=0.11$, $p=0.03$) and displayed no correlation whatsoever with the other dose-related values.

As visible in table 7, the VIF varied between 1.01 and 2.16 and was throughout far less than 5, which is regarded as threshold for the risk of multicollinearity. Hence, overall, no signs of multicollinearity were evident, which was underlined by rather low standard errors of the predictor coefficients (with exception to the factor level "concomittant+ adjuvant CTx" of the variable SEQUENCE).

Variable / Model	SMOKING	AGE	SEQUENCE	EUD	SUPRAGLOTTIC DMAX	GLOTTIC DMAX	SUPRAGLOTTIC DMEAN	GLOTTIC DMEAN
NL2a	1.06	1.31	1.53		1.23	1.29		
NL2b	1.02	1.02			1.19	1.2		
NL2c	1.01	1.02		1.01				
NL2d					1.19	1.19		
NL2e	1.10	1.10					2.11	2.1
NL2f	1.12	1.5	1.44				2.16	2.16
NL2g							2.08	2.08
NL1a	1.06	1.05			1.19	1.21		
ALL2a	1.16	1.27	1.3				1.27	1.21
ALL2b							1.05	1.05
L2a							1.02	1.02

Table 7. To assess whether multicollinearity of the variables included in the different models might be an issue, the VIF was calculated. Given that the VIF was less than 5 in all cases, no difficulties with multicollinearity were noticeable

3.5.3) Model development process

Model development was carried out with the training cohort data and started with a univariable regression analysis of the candidate predictors. Based on the results, the 5 variables BASELINE EDEMA, HPV, T-STAGE, HISTOLOGY, and TECHNIQUE were excluded from multivariable analysis given that they had a p-value >0.2. Detailed results from the univariable analysis are summarized in the appendix (A3).

Since new questions arose in the course of model development, a total of 11 models were set up and analyzed in detail. Table 8 provides an overview of the models and their components. To facilitate understanding, the models were given names reflecting the components of the respective model. The names were created according to the following scheme:

- NL = non-laryngeal sub-cohort
- L = laryngeal sub-cohort
- ALL = total cohort

- 1 = outcome laryngeal edema 1 or higher
- 2 = outcome laryngeal edema 2 or higher

- a = 1st model
- b = 2nd model etc.

For example, the model NL2a was developed in the non-laryngeal sub-cohort (NL), had the outcome "laryngeal edema 2 or higher" (2) and was the 1st model (a) that was set up (a). The following models were named according to this scheme. The rationale, background, and details of the different models are explained in detail in the following sections.

Model	Predictors	Cohort/Sub-cohort	Outcome
NL2a	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2b	SMOKING, AGE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2c	SMOKING, AGE, EUD	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2d	SUPRAGLOTTIC DMAX, GLOTTIC DMAX	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2e	SMOKING, AGE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2f	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL2g	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	Non-laryngeal HNC patients	laryngeal edema 2 or higher
NL1a	SMOKING, AGE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	Non-laryngeal HNC patients	laryngeal edema 1 or higher
ALL2a	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	All HNC patients	laryngeal edema 2 or higher
ALL2b	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	All HNC patients	laryngeal edema 2 or higher
L2a	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	Laryngeal HNC patients	laryngeal edema 2 or higher

Table 8. Overview of the 11 models with their different predictors, cohorts/sub-cohorts, and outcome definitions.

Model NL2a

The first model development was rendered without the candidate predictors GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN to avoid redundancy, since it was assumed that they are already represented by the laryngeal EUD. Therefore, the following 7 predictors were used in the model development process: EUD, SMOKING, SEQUENCE, LOCATION, AGE, SUPRAGLOTTIC DMAX, and GLOTTIC DMAX.

Like in Rancati's publications, model development was executed with the training sub-cohort consisting only of patients with HNCs outside the larynx. Analysis of the candidate predictors revealed that GLOTTIC DMAX and SUPRAGLOTTIC DMAX required transformation to approximate the normal distribution. For this purpose, a quadratic transformation of GLOTTIC DMAX and a cubic transformation of SUPRAGLOTTIC DMAX was executed. Laryngeal EUD calculation was carried out with all dose data available in the training data set. Stepwise regression led to an NTCP-model encompassing the 5 predictors SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMAX, and GLOTTIC DMAX, with the coefficients as given in table 9 (for reasons of clarity, the coefficients of the other models are shown here in advance).

	Variable	**Variable levels**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Model NL2a	(Intercept)		-45.58	22.26	0.04
	SMOKING	Yes	1.81	1.13	0.11
	SEQUENCE	concomittant CTx	1.46	0.72	0.04
		accelerated RT with cetuximab	2.09	0.7	0.00
		concomittant+ adjuvant CTx	-12.73	1381.64	0.99
	AGE		0.08	0.03	0.00
	SUPRAGLOTTIC DMAX		8.72	5.28	0.1
	GLOTTIC DMAX		-1.21	0.64	0.06
Model NL2b	(Intercept)		-42.5	21.41	0.05
	SMOKING	Yes	1.83	1.12	0.1
	AGE		0.05	0.02	0.02
	SUPRAGLOTTIC DMAX		8.86	5.11	0.08
	GLOTTIC DMAX		-1.62	0.61	0.01
Model NL2c	(Intercept)		-9.09	1.97	0.00
	EUD		0.39	0.17	0.02
	SMOKING	Yes	1.84	1.05	0.08
	AGE		0.05	0.02	0.02
Model NL2d	(Intercept)		-41.78	21.81	0.06
	SUPRAGLOTTIC DMAX		9.88	5.2	0.06
	GLOTTIC DMAX		-1.66	0.6	0.01

Model NL2e	(Intercept)		-16.72	3.21	0.00
	SMOKING	Yes	2.0	1.12	0.07
	AGE		0.06	0.02	0.01
	SUPRAGLOTTIC DMEAN		0.01	0.03	0.67
	GLOTTIC DMEAN		0.13	0.05	0.01
Model NL2f	(Intercept)		-17.9	3.49	0.00
	SMOKING	Yes	1.99	1.16	0.08
	SEQUENCE	concomittant CTx	1.31	0.72	0.07
		accelerated RT with cetuximab	1.84	0.69	0.01
		concomittant+ adjuvant CTx	-12.69	1364.2	0.99
	AGE		0.09	0.03	0.00
	SUPRAGLOTTIC DMEAN		0.11	0.05	0.03
	GLOTTIC DMEAN		0.01	0.03	0.78
Model NL2g	(Intercept)		-10.88	2.24	0.00
	SUPRAGLOTTIC DMEAN		0.12	0.05	0.01
	GLOTTIC DMEAN		0.02	0.03	0.49
Model NL1a	(Intercept)		-8.45	7.96	0.288
	SMOKING	Yes	-0.05	0.42	0.91
	AGE		-0.00	0.01	0.97
	SUPRAGLOTTIC DMEAN		2.28	1.88	0.23
	GLOTTIC DMEAN		-2.18	0.42	0.00
Model ALL2a	(Intercept)		-9.16	1.63	0.00
	SMOKING	Yes	0.11	0.44	0.81
	SEQUENCE	concomittant CTx	0.1	0.34	0.78
		accelerated RT with cetuximab	0.62	0.39	0.12
		concomittant+ adjuvant CTx	-12.88	545.44	0.98
	AGE		0.01	0.01	0.51
	SUPRAGLOTTIC DMEAN		0.03	0.01	0.01
	GLOTTIC DMEAN		0.07	0.02	0.00
Model ALL2b	(Intercept)		-8.63	1.21	0.00
	SUPRAGLOTTIC DMEAN		0.04	0.01	0.00
	GLOTTIC DMEAN		0.07	0.02	0.00
Model L2a	(Intercept)		-12.43	5.02	0.01
	SUPRAGLOTTIC DMEAN		0.03	0.01	0.03
	GLOTTIC DMEAN		0.13	0.07	0.08

Table 9. Overview of the 11 NTCP-models with their coefficients, the parameter estimates, and the associated standard errors and p-values.

After adjustment for optimism, the model NL2a had an AUC of 0.84 (95% CI: 0.81-0.86) and an R^2 of 0.19 (95% CI: 0.19-0.19). The corrected values for the calibration intercept and slope with their 95% confidence intervals were 0.05 (0.01-0.09) and 1.01 (1.0-1.02), respectively (table 10; to ensure clarity, the table also covers the values of the other models of the non-laryngeal cohort). Since the variable SEQUENCE was not present in the validation data set, no external validation of the model was accomplishable.

Model	**Predictors**	**cvAUC** (CI .025-CI .975)	**Intercept** (CI .025-CI .975)	**Slope** (CI .025-CI .975)	**R^2** (CI .025-CI .975)	**External Validation AUC** (CI .025-CI .975)
NL2a	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX	0.84 (0.81-0.86)	0.05 (0.01-0.09)	1.01 (1.0-1.02)	0.19 (0.19-0.19)	NA
NL2b	SMOKING, AGE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX	0.8 (0.78-0.83)	0.11 (0.07-0.15)	1.01 (1.0-1.01)	0.14 (0.14-0.14)	0.84 (0.71-0.97)
NL2c	SMOKING, AGE, EUD	0.69 (0.64-0.74)	0.08 (0.04-0.11)	1.01 (1.0-1.02)	0.14 (0.14-0.14)	0.69 (0.38-1.0)
NL2d	SUPRAGLOTTIC DMAX, GLOTTIC DMAX	0.79 (0.76-0.82)	0.07 (0.02-0.13)	1.01 (1.0-1.02)	0.14 (0.14-0.14)	0.76 (0.51-1.0)
NL2e	SMOKING, AGE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.84 (0.77-0.91)	0.05 (0.04-0.09)	1.01 (0.99-1.03)	0.22 (0.22-0.22)	0.97 (0.95-0.99)
NL2f	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.86 (0.74-0.97)	0.06 (0.04-0.08)	1.01 (1.0-1.01)	0.26 (0.26-0.26)	NA
NL2g	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.82 (0.80-0.85)	0.05 (0.01-0.08)	1.01 (1.0-1.01)	0.20 (0.20-0.20)	0.92 (0.88-0.96)

Table 10. Overview of adjusted performance measures (AUC, calibration intercept, calibration slope, R^2) with their confidence intervals for model NL2a-g, developed in the sub-cohort of non-laryngeal HNC patients. The outcome was defined as "laryngeal edema 2 or higher".

Model NL2b

To ascertain complete independence of the data sets, the validation cohort was not analyzed before the first model had been developed with the training data set. However, when attempting to externally validate model NL2a it became evident that the predictor SEQUENCE was not covered in the validation cohort. Hence, a new model (NL2b) was developed, in which SEQUENCE was omitted from the list of candidate predictors (in addition to GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN). Therefore, 6 predictors were available, and the stepwise regression resulted in a model encompassing the predictors SMOKING, AGE, SUPRAGLOTTIC DMAX, and GLOTTIC DMAX with the coefficients given in table 9. After adjustment for optimism, model NL2b had an AUC of 0.80 (95% CI: 0.78-0.83), an R^2 of 0.14 (95% CI: 0.14-0.14), a calibration intercept of 0.11 (0.07-0.15), and a calibration slope of 1.01 (1.0-1.01) (table 10).

The external validation of model NL2b resulted in an AUC of 0.84 (95% CI: 0.71-0.97) and an R^2 of 0.02 (95% CI: 0.01-0.02). The calibration intercept was -0.52 (95% CI: -0.82-(-0.03)) and the slope 0.88 (95% CI: 0.82-0.99).

Model NL2c

In the stepwise regression of model NL2b, the parameters SUPRAGLOTTIC DMAX and GLOTTIC DMAX were chosen as dose-related predictors, while the EUD was deleted during model development and hence was not represented in the final model. However, to assess the impact of the non-dose variables in direct comparison to Rancati's model (only EUD), model NL2c was set up manually, i.e., there was no stepwise regression, but the parameters SMOKING, AGE, and EUD were specifically selected. Thus, model NL2c differed from Rancati's model only in the non-dose-dependent variables SMOKING and AGE.

The EUD was calculated with all dose information contained in the training data set and a squared transformation of the EUD was executed to approximate the normal distribution. After cross-validation, model NL2c had an AUC of 0.69 (95% CI: 0.64-0.74) and an R^2 of 0.14 (95% CI: 0.14-0.14), which represented a considerable worse performance with wider CIs than model NL2b (table 10). The calibration was quantified with an intercept of 0.08 (0.04-0.11) and a slope of 1.01 (1.0-1.02).

External validation of model NL2c resulted in a cross-validated AUC of 0.69 (95% CI: 0.38-1.0) and an R^2 of 0.01 (95% CI: 0.01-0.01) with wide CIs due to the limited number of the outcome in the validation data set. The calibration intercept was calculated as -0.64 (95% CI: -0.87-(-0.36)) and the calibration slope as 0.92 (95% CI: 0.88-0.98).

Model NL2d

The juxtaposition of model NL2b and NL2c had shown that SUPRAGLOTTIC DMAX and GLOTTIC DMAX in combination with the non-dosimetric variables SMOKING and AGE led to a better model performance than EUD with SMOKING and AGE. Therefore, it was hypothesized that SUPRAGLOTTIC DMAX and GLOTTIC DMAX are more precise predictors than EUD when it comes to predicting post-radiogenic laryngeal edema. To test this hypothesis, model NL2d was manually built that contained only the predictors SUPRAGLOTTIC DMAX and GLOTTIC DMAX.

Model analysis showed an AUC of 0.79 (95% CI: 0.76-0.82) and an R^2 of 0.14 (95% CI: 0.14-0.14) (table 10). Hence, performance levels were superior to the ones estimated with Rancati's model, indicating that SUPRAGLOTTIC DMAX and GLOTTIC DMAX were indeed better suited representatives of the dose than the variable EUD. Quantification of the calibration gave an intercept of 0.07 (0.02-0.13) and a slope of 1.01 (1.0-1.02), respectively.

External validation of model NL2d resulted in an AUC of 0.76 (95% CI: 0.51-1.0), an R^2 of 0.01 (95% CI: 0.01-0.01), a calibration intercept of 1.81 (96% CI: -2.17-5.57) and a slope of 1.35 (95% CI: 0.44-1.86).

Model NL2e

From a clinical point of view, it was unexpected that representatives of the maximum dose applied (SUPRAGLOTTIC DMAX and GLOTTIC DMAX) contributed more to the model's predictive quality than the EUD, which incorporates both the irradiated volume and the average dose. To pursue this question further and examine the influence of the mean dose on model estimations, the parameters SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN were included in model NL2e as candidate predictors and a new NTCP model was developed. The variable SEQUENCE was removed from the list of

possible predictors to allow for external validation. Therefore, model development was exhibited with the candidate predictors EUD, SMOKING, LOCATION, AGE, SUPRAGLOTTIC DMAX, GLOTTIC DMAX, SUPRAGLOTTIC DMEAN, and GLOTTIC DMEAN.

Stepwise regression resulted in a model with the non-dosimetric predictors AGE and SMOKING, hence the same as included in model NL2b. Of notice, with respect to the dosimetric variables the combination of SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN were chosen for the final model instead of SUPRAGLOTTIC DMAX and GLOTTIC DMAX (table 9).

Evaluation of model NL2e demonstrated a corrected AUC of 0.84 (95% CI: 0.77-0.91) and an R^2 of 0.22 (95% CI: 0.22-0.22) as well as a calibration intercept of 0.05 (0.04-0.09) and a slope of 1.01 (0.99-1.03) (table 10). The external validation showed a cross-validated AUC of 0.97 (95% CI: 0.95-0.99), an R^2 of 0.01 (95% CI: 0.01-0.01), a calibration intercept of 0.07 (95% CI: -0.24-0.36), and a calibration slope of 1.02 (95% CI: 0.95-1.06).

Since the performance results were higher than those of the comparable model NL2b, the representatives of the mean dose, SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN, seemed to improve the precision of the model estimation when compared to SUPRAGLOTTIC DMAX and GLOTTIC DMAX.

Model NL2f

The juxtaposition of model NL2a and model NL2b had indicated the positive contribution of the variable SEQUENCE. Despite knowing that the parameter SEQUENCE was not recorded in the validation data set and hence, no external validation was possible, the predictor SEQUENCE was added to the list of candidate predictors and stepwise regression was applied with a total of 9 variables. The final model NL2f included the parameters SEQUENCE, SMOKING, AGE, GLOTTIC DMEAN, and SUPRAGLOTTIC DMEAN (table 9), which is equivalent to model NL2e with the addition of the predictor SEQUENCE. Assessment of model NL2f revealed a cross-validated AUC of 0.86 (95% CI: 0.74-0.97) and an R^2 of 0.26 (95% CI: 0.26-0.26), i.e., slightly better performance parameters than model NL2e (table 10). The

calibration parameters were quantified as an intercept of 0.06 (0.04-0.08) and a slope of 1.01 (1.0-1.01), respectively.

Model NL2g

The interim conclusion of the evaluation of model NL2a-f was that model NL2f containing the predictors SEQUENCE, SMOKING, AGE, GLOTTIC DMEAN, and SUPRAGLOTTIC DMEAN executed the best model performance so far with respect to estimating the risk of post-radiogenic laryngeal edema in non-laryngeal HNC patients. Moreover, the comparison of model NL2b and model NL2c indicated that SUPRAGLOTTIC DMAX and GLOTTIC DMAX were more suitable dose-related predictors than the laryngeal EUD. Additionally, the juxtaposition of model NL2f and model NL2a suggested that the combination of SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN outperformed SUPRAGLOTTIC DMAX and GLOTTIC DMAX. Thus, regarding the dose-related predictors the following ranking, starting from the best, was assumed: GLOTTIC DMEAN and SUPRAGLOTTIC DMEAN > SUPRAGLOTTIC DMAX and GLOTTIC DMAX > EUD. To prove this hypothesis, model NL2g was manually built, which solely contained the dose-relevant parameters SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN (table 9). This model should be set against model NL2d, which included only the parameters SUPRAGLOTTIC DMAX and GLOTTIC DMAX, to verify the proposed hypothesis.

Model NL2g had a corrected AUC of 0.82 (95% CI: 0.80-0.85) with an R^2 of 0.20 (95% CI: 0.2-0.2) and hence surpassed model NL2d (AUC=0.79; 95% CI: 0.76-0.82), but performed worse than model NL2e and NL2f (table 10). Quantification of the calibration resulted in an intercept of 0.05 (0.01-0.08) and a slope of 1.01 (1.0-1.01). External validation of model NL2g revealed a corrected AUC of 0.92 (95% CI: 0.88-0.96), an R^2 of 0.02 (95% CI: 0.01-0.02), a calibration intercept of -0.37 (95%:-0.66-(-0.09)), and a calibration slope of 0.93 (95% CI: 0.86-0.97).

Based on these results, it was concluded that the clinical predictors SMOKING and AGE significantly enhanced model performance, which was even better when supplemented with the variable SEQUENCE. With regard to the dose-dependent variables, it was confirmed that the combination of SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN increased the precision of the model estimation best, while the EUD

showed the lowest contribution in this regard. Since the EUD was calculated from the DVH and thus the irradiated volume, it followed that the influence of the irradiation volume was apparently less decisive for the presence of post-radiogenic diseases than previously thought.

Model NL1a

To ground the external validation on a higher number of patients with the disease of interest, the outcome was redefined to "laryngeal edema grade 1 or higher", even though grade 1 denotes an asymptomatic edema with assumingly limited clinical consequence. Due to the redefinition, the number of patients with the outcome rose in the training data set from 26 to 76 (on average) and in the validation data set from 2 to 28, respectively, thereby enhancing the stability of the calculations.

Similar to the other models, an external validation of Rancati's model with the newly defined outcome was first carried out. Computation revealed an AUC of 0.56 (95% CI: 0.53-0.60), an R^2 of 0.05 and calibration parameters with an intercept of -1.1 (95% CI: -5.2-(-0.02)) and a slope of 1.34 (95% CI: -1.2-4.2).

If models including the variable SEQUENCE were left out, model NL2e with the parameters SMOKING, AGE, SUPRAGLOTTIC DMEAM, and GLOTTIC DMEAN showed the best estimation performance so far. Equivalent to this, the new model NL1a was created manually and the model performance was analyzed taking into account the new outcome. Assessment of model NL1a displayed a corrected AUC of 0.76 (95% CI: 0.75-0.76), a cross-validated R^2 of 0.19 (95% CI: 0.19-0.19), a calibration intercept of 0.04 (0.03-0.04), and a slope of 1.01 (1.01-1.01) (Table 11). Obviously, the significantly larger number of outcomes had led to a stabilization of the predictions, as mirrored by small CIs. External validation of model NL1a yielded a corrected AUC of 0.72 (95% CI: 0.63-0.82), an R^2 of 0.11 (95% CI: 0.11-0.13), a calibration intercept of -0.08 (95% CI: -0.41-0.1) and a calibration slope of 0.97 (95% CI: 0.91-1.01)

Model	Predictors	cvAUC (CI .025-CI .975)	Intercept (CI .025-CI .975)	Slope (CI .025-CI .975)	R^2 (CI .025-CI .975)	External Validation AUC (CI .025-CI .975)
NL1a	SMOKING, AGE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.76 (0.75-0.76)	0.04 (0.03-0.04)	1.01 (1.01-1.01)	0.19 (0.19-0.19)	0.83 (0.63-0.82)

Table 11. Summary of the adjusted performance measures (AUC, calibration intercept, calibration slope, R^2) with their confidence intervals of model NL1a, which was analyzed in the non-laryngeal subcohort. As indicated by the abbreviation, in this model the definition of the outcome was different ("laryngeal edema 1 or higher"), when compared to model NL2a-g.

Model ALL2a

To explore to what extent model estimation was altered by modifications of the cohort, namely when all HNC patients were included in model analysis, model ALL2a was set up, which contained the so far best-suited predictors SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN without performing a stepwise regression beforehand. For the calculations, the definition of the outcome was reset to the original ("laryngeal edema 2 or higher"). Of the 750 patients of the total cohort, n=74 participants were diagnosed with the outcome at the index date.

The external validation of Rancati's LOGEUD model with the total cohort yielded an AUC of 0.61 (95% CI: 0.59-0.62), an R^2 of 0.02 (95% CI: 0.01-0.03), a calibration intercept of -0.48 (95% CI: -10.3-9,34), and a calibration slope of 1.3 (95% CI: -2.34-3.15).

The examination of model ALL2a resulted in a corrected AUC of 0.71 (95% CI: 0.70-0.72), an R^2 of 0.11 (95% CI: 0.11-0.11), a calibration intercept of 0.05 (0.04-0.06), and a slope of 1.02 (1.02-1.02) (table 12). Given that the variable SEQUENCE was not included in model ALL2a, it could not be externally validated with the validation cohort.

Nevertheless, analysis demonstrated that, as expected, the modification of the cohort has altered model performance.

Model ALL2b

The rather moderate estimation performance of model ALL2a led to efforts to develop a new, possibly more suitable model in the total cohort. Stepwise regression considering all 9 candidate predictors resulted in a model consisting solely of the variables SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN. Analysis showed a corrected AUC of 0.75 (95% CI: 0.74-0.77) and a cross-validated R^2 of 0.14 (95% CI: 0.14-0.14). The calibration intercept was 0.04 (95% CI: 0.03-0.05), and the slope 1.02 (95% CI: 1.02-1.02) (table 12). Of notice, model ALL2b exhibited an improved model performance in contrast to model ALL2a, which additionally involved the non-dosimetric parameters AGE, SMOKING, and SEQUENCE.

External validation of the model revealed an AUC of 0.78 (95% CI: 0.67–0.89), an R^2 of 0.12 (95% CI: 0.11-0.13), a calibration intercept of 0.28 (95% CI: -0.11-0.77), and a calibration slope of 1.07 (95% CI: 0.31-1.69).

Model	**Predictors**	**cvAUC** (CI .025-CI .975)	**Intercept** (CI .025-CI .975)	**Slope** (CI .025-CI .975)	**R^2** (CI .025-CI .975)	**External Validation AUC** (CI .025-CI .975)
ALL2a	SMOKING, AGE, SEQUENCE, SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.71 (0.7-0.72)	0.05 (0.04-0.06)	1.02 (1.02-1.02)	0.11 (0.11-0.11)	NA
ALL2b	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.75 (0.74-0.77)	0.04 (0.03-0.05)	1.02 (1.02-1.02)	0.14 (0.14-0.14)	0.78 (0.67–0.89)

Table 12. Performance measures of model ALL2a and ALL2b, which have been estimated in the total training cohort with the original outcome ("laryngeal edema 2 or higher"). Model ALL2b with solely dosimetric variables outperformed model ALL2a, which contained non-dosimetric factors, too.

Model L2a

To cover all variants in the composition of the study cohort, a model development was finally carried out in the sub-cohort that only contained patients with a primary tumor located in the larynx. This laryngeal sub-cohort of the training data set encompassed n=334 patients, with n=49 (14.7%) displaying laryngeal edema grade 2 or more at the index date. External validation of Rancati's model with the laryngeal sub-cohort demonstrated an AUC of 0.51 (95% CI: 0.42-0.60), an R^2 of 0.04 (95% CI: 0.03-0.04) and the calibration was quantified with an intercept of -0.8 (95% CI: -4.1-1.83) and a slope of 1.53 (95% CI: -0.83-2.92). Stepwise regression yielded a new model that contained only the dosimetric values SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN. Hence, similar to model ALL2b, no non-dosimetric variables were part of the final model. Examination of model L2a displayed a corrected AUC of 0.66

(95% CI: 0.58-0.74), a cross-validated R^2 of 0.06 (0.05-0.06), a calibration intercept of 0.24 (-0.19-0.46), and a slope of 1.17 (1.02-1.27) (table 13).

The external validation of model L2a revealed an AUC of 0.74 (95% CI: 0.61-0.86), an R^2 of 0.03 (95% CI: 0.03-0.04), a calibration intercept of 0.7 (95% CI: -2.25-2.3), and a calibration slope of 1.39 (95% CI: 0.14-2.07).

Model	**Predictors**	**cvAUC** (CI .025-CI .975)	**Intercept** (CI .025-CI .975)	**Slope** (CI .025-CI .975)	**R^2** (CI .025-CI .975)	**External Validation AUC** (CI .025-CI .975)
L2a	SUPRAGLOTTIC DMEAN, GLOTTIC DMEAN	0.66 (0.58-0.74)	0.24 (-0.19-0.46)	1.17 (1.02-1.27)	0.06 (0.05-0.06)	0.74 (0.61-0.86)

Table 13. Model L2a was the only model that was analyzed in the sub-cohort of laryngeal HNC patients, applying the original outcome definition of "laryngeal edema 2 or higher".

3.5.4) Overview of the performance of the 11 NTCP-models

In summary, the evaluations of the 11 models resulted in the following findings:
I.) Of the dose-related variables, the combination of SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN improved model performance the most, followed by SUPRAGLOTTIC DMAX and GLOTTIC DMAX. As opposed to this, laryngeal EUD provided the least positive input to the model estimates.
II.) In the sub-cohort of non-laryngeal HNC patients, the best model estimate was achieved with a combination of dosimetric and non-dosimetric predictors. The variable SEQUENCE strengthened model performance.
III.) In the total cohort and the sub-cohort of laryngeal HNC patients, the model that was based solely on dosimetric variables accomplished the best estimates and non-dosimetric parameters did not significantly contribute to model performance.

3.5.5) Comparison of external validation

Rancati's LOGEUD model and all new NTCP-models without the parameter SEQUENCE were externally validated with the same independent validation data set. The comparison of the results from the external validation of model NL2b, NL2c, NL2d, and NL2g, respectively, as opposed to Rancati's model, showed that the performance of the new NTCP-models was significantly better. While the discrimination power, quantified by the AUC, was only 0.51 in the external validation of Rancati's model, the AUC lay between 0.69-0.97 for the NTCP-models. The highest AUC as well as a good calibration was found in model NL2e, which contained the variables SMOKING, AGE, SUPRAGLOTTIC DMEAN, and GLOTTIC DMEAN. These findings underlined the advantages of models relying on both dosimetric and non-dosimetric variables when compared to models based on the EUD alone.

4.) Discussion

Principal findings

Laryngeal edema is a potential long-term complication of RT of HNC and may have significant consequences for the patient's quality of life. NTCP-models can help to strike a balance between optimized protection of the OARs without compromising tumor control. The main aim of the present study was the development of new NTCP-models to predict post-radiogenic laryngeal edema in HNC patients and the principal findings were:

(I) The external validation of Rancati's LOGEUD model (8,62), which relies solely on dosimetric variables, yielded a moderate discriminative power. Probably, this is at least partly explained by differences in the patient cohorts with a prevalence of the outcome in non-laryngeal HNC patients of 52% in Rancati's cohort (8) compared to only 7% in the current study patients.

(II) During model development in the non-laryngeal HNC cohort, model NL2f including non-dosimetric (AGE, SMOKING, SEQUENCE) and dosimetric predictors (SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN) was best and displayed good calibration and discrimination performance with narrow confidence intervals.

(III) The benefit of the combination of dosimetric and non-dosimetric variables was demonstrated using the independent validation cohort with a significantly higher AUC of the models composed of dosimetric and non-dosimetric variables compared to Rancati's LOGEUD model.

(IV) Within the dosimetric variables, the combination of SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN achieved better results than SUPRAGLOTTIC DMAX and GLOTTIC DMAX, which in turn outperformed the EUD.

(V) Modifications in the composition of the cohort or the outcome definition required adjustments of the relevant predictors.

(VI) Despite high correlation between the dose-related variables, multicollinearity was not evident in the present study as underlined by a VIF of 1.01-2.16.

Study strengths

The greatest strength of the present study was the very large study population as well as the additional, independent patient cohort for external validation, in which time-

shifted data collection was executed. In addition to the large number of patients, the high quantity of variables that were acquired from all study participants at different times, provided very useful supplementary information. Furthermore, the present study impresses with comprehensive model developments and updates to illuminate the scientific question from all sides, e.g., the impact of modifications of the outcome or the (sub-)cohort on model performance.

Study weaknesses

The following caveats in relation to the presents study results need to be considered. First, the outcome was at least partly subjective and influenced by interobserver variability, which may have hampered the quality and performance of the NTCP-models (63). Even if the assessment of the presence of laryngeal edema was carried out in a standardized manner employing a fiber-optic examination and using defined criteria of the LENT-SOMA classification, the judgment lay ultimately in the responsibility of the examiner and was not tied to hard facts. In addition, the patient's subjective assessment probably also played a role in determining whether an edema was classified as symptomatic or asymptomatic. Second, the extensive evaluations and comparisons did not involve an independent model development process in each of the 11 new NTCP-models. However, it was carried out in the majority of models, namely, in model NL2a, NL2b, NL2e, NL2f, ALL2b, and L2a, respectively. Therefore, in each of the three (sub-)cohorts (non-laryngeal subcohort, laryngeal subcohort, total cohort) an independent model development was conducted at least once. Third, despite the great endeavor and effort involved in the study, the data sets included missing values. To compensate for this shortcoming, multiple imputations were made, as recommended in the literature for such a situation. Forth, despite the large number of study participants, the 4 criteria proposed by Riley could not be fully met (40). This is partly due to the rarity of the outcome, which is a frequently encountered drawback in the development process of prediction models in radio-oncology. The rare occurrence of the outcome is largely attributable to medical progress regarding early disease detection as well as technological advancements, providing less invasive therapeutic options.

Comparison with other studies

Compared to previously published studies on NTCP-prediction models of post-radiogenic laryngeal edema in HNC patients, model developments were carried out with by far the largest number of patients, even after almost halving of the training cohort when the laryngeal carcinoma patients were excluded. The main evaluations were accomplished with a total of n=416 patients, which considerably surpasses Rancati's (n=48) (8) and Jakobi's study (n=45) (34,35).

The low occurrence of the outcome was predominantly encountered in the external validation of the new NTCP-models. Given that performance estimates in the validation cohort were grounded on only 2 outcomes in the sub-cohort of non-laryngeal HNC patients, interpretation of the results was only possible to a limited extent. Because of this, the definition of the outcome in model NL1a was changed, although a grade 1, i.e., asymptomatic edema, probably owns only little clinical relevance. Due to this alteration, the prevalence of the outcome increased to 19% in the training data set and to 16% in the validation data set. This enabled a more stable and reliable estimation of model performance during the external validation, which is important to ensure generalizability and clinical use of the validated NTCP-models (46,63–66).

Not only Bahn and Alber but also some other authors generally question the suitability of the ROC curve analysis for NTCP-models (67–69). They argue that NTCP-models are continuous, and the AUC has been formulated for binary classifiers. However, the outcome in the present study was binary and relied on the judgment according to LENT-SOMA. Because of this, the ROC curve analysis deemed suitable in the current study.

The 5 variables HPV, T-STAGE, HISTOLOGY, BASELINE EDEMA, and TECHNIQUE were excluded from model development after univariable analysis. Regarding HPV, the reason for this likely is the small amount of HPV data in the training data set (6%), since routine testing for HPV only began a few years ago, as mirrored in the higher HPV data availability in the validation data set (86%), which was acquired later in time. Therefore, examination of HPV in the context of NTCP-models should be repeated in future studies. T-STAGE was assumably not significant in model development as 76% of patients experienced an advanced tumor stage (Stage IVa or IVb) and therefore rather little variation within the cohort was present. The variable

HISTOLOGY was chosen as surrogate for pathological differentiation, which had been used as predictor by Bae et al. (49). However, as expected in HNCs, SCC was by far the most common tumor cell type, so there was almost no variation within the training cohort and the predictor was removed, consequently. Putative differences in pathophysiology are likely the rationale why the variable BASELINE LARYNGEAL EDEMA did not strengthen the predictive model for chronic edema. A potential reason for the removal of TECHNIQUE might be that the majority of patients in the training data set underwent IMRT therapy without much deviation within the cohort. Until about 2008, IMRT was the dominant RT technique in clinical use and thereafter has been increasingly replaced by VMAT. The trend towards growing employment of VMAT was mirrored in the juxtaposition of the training and the validation cohort. While almost three quarters of the patients in the training data set were treated with IMRT and only 16% got VMAT therapy, the latter was almost exclusively applied in the validation cohort. Despite the higher prevalence of laryngeal edema of 14.7% in the laryngeal sub-cohort compared to the other sub-cohorts, no distinct effect of the parameter LOCATION was visible and therefore the variable was excluded.

As opposed to the aforementioned variables, SEQUENCE was capable to significantly enhance the model estimates, which confirmed the importance of the multiple effects of CTx on tumor growth. Induction CTx can reduce the volume of the primary tumor before the start of RT, potentially enabling that both the total dose as well as the irradiation field may be confined. Concomitant CTx or treatment with paclitaxel elevates the radiation sensitivity of the tumor, presumably allowing a reduction of the radiation dose, too. Regarding the occurrence of post-radiogenic laryngeal edema, however, it needs to be considered that CTx always has an effect on the whole-body and is not limited to a certain part of the body. Hence, not only tumor cells but also healthy tissue becomes more radio-sensitive and thus might easier be harmed. As a result, patients with concomitant CTx inherit a greater risk of damage to adjacent OARs even at lower doses. This is underlined by the variable SEQUENCE: if the parameter is present in the model, the model estimate is better, i.e., SEQUENCE has an impact on the presence of the post-radiogenic complication.

A potential rationale, why AGE only stayed in the model of the non-laryngeal HNC cohort, might be that the laryngeal HNC patients made up 45% of the total cohort,

which is why characteristics typical of them had a distinct impact on the effect of this variable in the total cohort. As known from literature, laryngeal carcinoma patients are mostly diagnosed in the sixth or seventh decades of life (18). This was reflected in the laryngeal HNC sub-cohort, in which the mean age at the time of diagnosis was 65.1 years. No significant discrepancy in age between patients with and without post-radiogenic laryngeal edema was evident ($p=0.13$) and the variable AGE was removed from the prediction model in laryngeal carcinoma patients. In comparison, the group of non-laryngeal HNC patients was slightly younger, with a mean age of 61.5. Evaluation of their age distribution revealed that patients with laryngeal edema were significantly older than patients without edema (65.5 versus 61.2; $p=0.02$). This is also the rationale why the variable AGE led to a model improvement only in the sub-cohort of non-laryngeal HNC patients, while it was removed from the model in the laryngeal HNC patients and the total cohort. Since the age of non-laryngeal patients with edema corresponded approximately to the mean age of the laryngeal HNC patients with and without edema, the effect of the parameter AGE was reduced due to the mingling of the sub-cohorts.

Closer examination of the non-laryngeal HNC patients displayed that the ratio of smokers to non-smokers with and without laryngeal edema was significantly different ($p<0.01$). In participants without confirmed laryngeal edema, the smoker- vs. non-smoker-ratio was 5.7, while it reached 28.0 in patients suffering from edema. As opposed to that, no relevant discrepancies were found in the laryngeal HNC sub-cohort with a ratio of 8.8 in patients without edema and a comparable ratio of 7.2 in participants diagnosed with laryngeal edema, respectively. Even though, slight differences in the smoker-to-non-smoker ratio in patients with (6.7) vs. without laryngeal edema (10.1) were seen in the total cohort, obviously, this did not have a significant impact on the model estimates. Hence, the variable SMOKING was deleted from the list of predictors in the total as well as the laryngeal sub-cohort but was retained in the non-laryngeal cohort.

Apparently, the predictors SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN, i.e., the average dose hitting the larynx, had a greater impact on the presence of post-radiogenic laryngeal edema than the EUD which additionally incorporated the irradiated volume. However, in discussions conducted prior to the begin of the study,

it was hypothesized that the irradiated volume would play a distinct role when it comes to the development of a laryngeal edema: a larger irradiation field often requires a higher total dose and at the same time there is an increased chance that the larynx will also be affected during the irradiation. Yet, the review of the literature showed that similar results have already been described in NTCP-models for the prediction of radiation-induced hypothyroidism (70), in which the irradiated volume neither had a relevant influence on the occurrence of the post-radiogenic disease. It can be speculated whether an explanation might be that the information on the 3D-dose distribution is lost in DVH tools that condense the spatial expansion of the dose into a one-dimensional function (71). This could reduce the information content and thus the influence on an NTCP-model.

Even though, it is currently unknown to what extent prediction models for RT with photons are generalizable and transferable to proton-based RT, model NL2f constitutes a good starting point for setting up NTCP-models in proton therapy. Some authors have already made a commitment in this regard and are of the opinion that the NTCP-models developed for photon therapy are also valid for proton-based RT (72). Before clinical use, however, these assumptions should be checked in separate studies.

5.) Conclusion

In summary, the present study was able to demonstrate in a large cohort of non-laryngeal HNC patients that an NTCP-model including the predictors AGE, SMOKING, SEQUENCE, SUPRAGLOTTIC DMEAN and GLOTTIC DMEAN was most suitable for predicting post-radiogenic laryngeal edema. Thus, besides dose variables, non-dosimetric parameters played a distinct role, representing an important step towards individualized patient treatment. Probably related to the rare occurrence of the outcome, rather small changes in the composition of the cohort required adaptations of the relevant predictors. Hence, it will be the task of future studies to develop various NTCP-models to estimate the risk of the most critical post-radiogenic complications in different patient settings. The next step then would be to further externally validate and improve these models, so that in the long term they can enable a comparison of possible advantages of proton versus photon therapy regarding the risk of post-radiogenic complications (73). This would be another crucial step to optimize treatment of HNC patients in terms of reducing post-radiogenic complications without loss of tumor control.

6.) References

1. Cancer of the Oral Cavity and Pharynx - Cancer Stat Facts [Internet]. SEER. [cited 2022 Mar 13]. Available from: https://seer.cancer.gov/statfacts/html/oralcav.html
2. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009 Feb 23;338:b375.
3. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009 Mar 31;338:b604.
4. Van den Bosch L, Schuit E, van der Laan HP, Reitsma JB, Moons KGM, Steenbakkers RJHM, et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 Jul;148:151–6.
5. Porcher R, Jacot J, Wunder JS, Biau DJ. Identifying treatment responders using counterfactual modeling and potential outcomes. Stat Methods Med Res. 2019 Nov;28(10–11):3346–62.
6. Tambas M, Steenbakkers RJHM, van der Laan HP, Wolters AM, Kierkels RGJ, Scandurra D, et al. First experience with model-based selection of head and neck cancer patients for proton therapy. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 Oct;151:206–13.
7. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2013 Jun;107(3):267–73.
8. Rancati T, Fiorino C, Sanguineti G. NTCP modeling of subacute/late laryngeal edema scored by fiberoptic examination. Int J Radiat Oncol Biol Phys. 2009 Nov 1;75(3):915–23.
9. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49.
10. Auguste A, Deloumeaux J, Joachim C, Gaete S, Michineau L, Herrmann-Storck C, et al. Joint effect of tobacco, alcohol, and oral HPV infection on head and neck cancer risk in the French West Indies. Cancer Med. 2020 Sep;9(18):6854–63.

11. Jiang S, Dong Y. Human papillomavirus and oral squamous cell carcinoma: A review of HPV-positive oral squamous cell carcinoma and possible strategies for future. Curr Probl Cancer. 2017 Oct;41(5):323–7.
12. Steuer CE, El-Deiry M, Parks JR, Higgins KA, Saba NF. An update on larynx cancer. CA Cancer J Clin. 2017;67(1):31–50.
13. NKR Cijfers [Internet]. [cited 2022 Feb 15]. Available from: https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=526&fs%7Ctumor_id=78&fs%7Cregio_id=550&fs%7Cperiode_id=564%2C565%2C566%2C567%2C568%2C569%2C570%2C571%2C572%2C573%2C574%2C575%2C576%2C577%2C578%2C579%2C580%2C581%2C582%2C583%2C584%2C585%2C586%2C587%2C588%2C589%2C590%2C591%2C592%2C593%2C563%2C562%2C561&fs%7Cgeslacht_id=644&fs%7Cleeftijdsgroep_id=677&fs%7Cjaren_na_diagnose_id=687&fs%7Ceenheid_id=703&cs%7Ctype=line&cs%7CxAxis=periode_id&cs%7Cseries=epidemiologie_id&ts%7CrowDimensions=periode_id&ts%7CcolumnDimensions=&lang%7Clanguage=nl
14. Kuper H, Boffetta P, Adami HO. Tobacco use and cancer causation: association by tumour type. J Intern Med. 2002 Sep;252(3):206–24.
15. Boffetta P, Hashibe M. Alcohol and cancer. Lancet Oncol. 2006 Feb;7(2):149–56.
16. Mehanna H, Paleri V, West CML, Nutting C. Head and neck cancer--Part 1: Epidemiology, presentation, and prevention. BMJ. 2010 Sep 20;341:c4684.
17. German Society of Cancer. Evidence-based Guideline Laryngeal Cancer (Version 1.1, 2019) [S3-Leitlinie Diagnostik, Therapie und Nachsorge des Larynxkarzinoms] [Internet]. [cited 2021 Nov 14]. Available from: https://www.leitlinienprogramm-onkologie.de/leitlinien/larynxkarzinom/
18. Chatelet F, Wagner I, Bizard A, Hans S, Chabolle F, Bach CA. Does advanced age affect treatment of early glottic carcinoma? Eur Ann Otorhinolaryngol Head Neck Dis. 2021 Mar;138(2):68–72.
19. Murphy BA, Gilbert J, Ridner SH. Systemic and global toxicities of head and neck treatment. Expert Rev Anticancer Ther. 2007 Jul;7(7):1043–53.

20. Langendijk JA, Doornaert P, Verdonck-de Leeuw IM, Leemans CR, Aaronson NK, Slotman BJ. Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Aug 1;26(22):3770–6.

21. Earl MA, Shepard DM, Naqvi S, Li XA, Yu CX. Inverse planning for intensity-modulated arc therapy using direct aperture optimization. Phys Med Biol. 2003 Apr 21;48(8):1075–89.

22. Ouyang Z, Liu Shen Z, Murray E, Kolar M, LaHurd D, Yu N, et al. Evaluation of auto-planning in IMRT and VMAT for head and neck cancer. J Appl Clin Med Phys. 2019 Jul;20(7):39–47.

23. Leung WS, Wu VWC, Liu CYW, Cheng ACK. A dosimetric comparison of the use of equally spaced beam (ESB), beam angle optimization (BAO), and volumetric modulated arc therapy (VMAT) in head and neck cancers treated by intensity modulated radiotherapy. J Appl Clin Med Phys. 2019 Oct 8;20(11):121–30.

24. Becker W, Naumann HH, Pfaltz CR. Hals-Nasen-Ohrenheilkunde. Stuttgart New York: Thieme; 1983.

25. Scallan J, Huxley VH, Korthuis RJ. Pathophysiology of Edema Formation [Internet]. Capillary Fluid Exchange: Regulation, Functions, and Pathology. Morgan & Claypool Life Sciences; 2010 [cited 2022 Feb 27]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK53445/

26. Collins GS, Reitsma JB, Altman DG, Moons KGM, members of the TRIPOD group. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol. 2015 Jun;67(6):1142–51.

27. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1-73.

28. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016 Jan;4(2):30.

29. Radiotherapie bij hoofd-halstumoren - Richtlijn - Richtlijnendatabase [Internet]. [cited 2022 Feb 15]. Available from: https://richtlijnendatabase.nl/richtlijn/hoofd-halstumoren/radiotherapie_bij_hoofd-halstumoren.html
30. Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, 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. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2015 Oct;117(1):83–90.
31. LENT SOMA tables. Radiother Oncol J Eur Soc Ther Radiol Oncol. 1995 Apr;35(1):17–60.
32. Common Terminology Criteria for Adverse Events (CTCAE). 2017;147.
33. de Veij Mestdagh PD, Janssen T, Lamers E, Carbaat C, Hamming-Vrieze O, Vogel WV, et al. SPECT/CT-guided elective nodal irradiation for head and neck cancer: Estimation of clinical benefits using NTCP models. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2019 Jan;130:18–24.
34. Jakobi A, Bandurska-Luque A, Stützer K, Haase R, Löck S, Wack LJ, et al. Identification of Patient Benefit From Proton Therapy for Advanced Head and Neck Cancer Patients Based on Individual and Subgroup Normal Tissue Complication Probability Analysis. Int J Radiat Oncol Biol Phys. 2015 Aug 1;92(5):1165–74.
35. Jakobi A, Stützer K, Bandurska-Luque A, Löck S, Haase R, Wack LJ, et al. NTCP reduction for advanced head and neck cancer patients using proton therapy for complete or sequential boost treatment versus photon therapy. Acta Oncol Stockh Swed. 2015;54(9):1658–64.
36. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. Med Phys. 1997 Jan;24(1):103–10.
37. Henríquez FC, Castrillón SV. A quality index for equivalent uniform dose. J Med Phys Assoc Med Phys India. 2011;36(3):126–32.
38. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology. 2010 Jan;21(1):128–38.
39. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003 Sep;56(9):826–32.

40. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;368:m441.
41. Ensor J. Package 'pmsampsize' [Internet]. 2021 [cited 2021 Dec 30]. Available from: https://cran.r-project.org/web/packages/pmsampsize/pmsampsize.pdf
42. Jeans C, Brown B, Ward EC, Vertigan AE, Pigott AE, Nixon JL, et al. Comparing the prevalence, location, and severity of head and neck lymphedema after postoperative radiotherapy for oral cavity cancers and definitive chemoradiotherapy for oropharyngeal, laryngeal, and hypopharyngeal cancers. Head Neck. 2020 Nov;42(11):3364–74.
43. Bosch LV den, Laan HP van der, Schaaf A van der, Oosting SF, Halmos GB, Witjes MJH, et al. Patient-Reported Toxicity and Quality-of-Life Profiles in Patients With Head and Neck Cancer Treated With Definitive Radiation Therapy or Chemoradiation. Int J Radiat Oncol Biol Phys. 2021 Oct 1;111(2):456–67.
44. El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006 Mar 15;64(4):1275–86.
45. Cella L, Liuzzi R, Conson M, D'Avino V, Salvatore M, Pacelli R. Development of multivariate NTCP models for radiation-induced hypothyroidism: a comparative analysis. Radiat Oncol Lond Engl. 2012 Dec 27;7:224.
46. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1-73.
47. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001 Aug 1;54(8):774–81.
48. Rubin D. Multiple Imputation for Nonresponse in Surveys. In: Multiple Imputation for Nonresponse in Surveys [Internet]. John Wiley & Sons, Ltd; 1987 [cited 2022 Feb 28]. p. i–xxix. Available from: http://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316696.fmatter

49. Bae JS, Roh JL, Lee S wook, Kim SB, Kim JS, Lee JH, et al. Laryngeal edema after radiotherapy in patients with squamous cell carcinomas of the larynx and hypopharynx. Oral Oncol. 2012 Sep;48(9):853–8.
50. Machtay M, Moughan J, Farach A, Martin-O'Meara E, Galvin J, Garden AS, et al. Hypopharyngeal dose is associated with severe late toxicity in locally advanced head-and-neck cancer: an RTOG analysis. Int J Radiat Oncol Biol Phys. 2012 Nov 15;84(4):983–9.
51. Thomas M, Defraene G, Lambrecht M, Deng W, Moons J, Nafteux P, et al. NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2019 Dec;141:33–40.
52. El-Naggar: WHO classification of head and neck tumours - Google Scholar [Internet]. [cited 2022 Feb 27]. Available from: https://scholar.google.com/scholar_lookup?title=WHO+classification+of+head+and+neck+Tumours&author=AK+El-Naggar&author=JKC+Chan&author=JR+Grandis&author=T+Takata&author=PJ+Slootweg&publication_year=2017&
53. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chen C, Curado MP, et al. Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. J Natl Cancer Inst. 2007 May 16;99(10):777–89.
54. Caglar HB, Tishler RB, Othus M, Burke E, Li Y, Goguen L, et al. Dose to larynx predicts for swallowing complications after intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys. 2008 Nov 15;72(4):1110–8.
55. Nguyen NP, Abraham D, Desai A, Betz M, Davis R, Sroka T, et al. Impact of image-guided radiotherapy to reduce laryngeal edema following treatment for non-laryngeal and non-hypopharyngeal head and neck cancers. Oral Oncol. 2011 Sep;47(9):900–4.
56. Sopori M. Effects of cigarette smoke on the immune system. Nat Rev Immunol. 2002 May;2(5):372–7.
57. Harrison LB, Chadha M, Hill RJ, Hu K, Shasha D. Impact of tumor hypoxia and anemia on radiation therapy outcomes. The Oncologist. 2002;7(6):492–508.

58. Lerman J, Hennequin C, Etienney I, Abramowitz L, Goujon G, Gornet JM, et al. Impact of tobacco smoking on the patient's outcome after (chemo)radiotherapy for anal cancer. Eur J Cancer. 2020 Dec 1;141:143–51.
59. Majchrzak E, Szybiak B, Wegner A, Pienkowski P, Pazdrowski J, Luczewski L, et al. Oral cavity and oropharyngeal squamous cell carcinoma in young adults: a review of the literature. Radiol Oncol. 2014 Jan 22;48(1):1–10.
60. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, et al. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. J Natl Cancer Inst. 2000 May 3;92(9):709–20.
61. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec;49(12):1373–9.
62. Rancati T, Schwarz M, Allen AM, Feng F, Popovtzer A, Mittal B, et al. Radiation dose-volume effects in the larynx and pharynx. Int J Radiat Oncol Biol Phys. 2010 Mar 1;76(3 Suppl):S64-69.
63. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019 Jan 1;170(1):W1–33.
64. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiother Oncol J Eur Soc Ther Radiol Oncol. 2018 Jun;127(3):370–3.
65. Sharabiani M, Clementel E, Andratschke N, Hurkmans C. Generalizability assessment of head and neck cancer NTCP models based on the TRIPOD criteria. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 May;146:143–50.
66. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009 May 28;338:b605.
67. Bahn E, Alber M. On the limitations of the area under the ROC curve for NTCP modelling. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 Mar;144:148–51.
68. Byrne S. A note on the use of empirical AUC for evaluating probabilistic forecasts. Electron J Stat. 2016 Jan;10(1):380–93.
69. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007 Feb 20;115(7):928–35.

70. Luo R, Wu VWC, He B, Gao X, Xu Z, Wang D, et al. Development of a normal tissue complication probability (NTCP) model for radiation-induced hypothyroidism in nasopharyngeal carcinoma patients. BMC Cancer. 2018 May 18;18(1):575.
71. Palma G, Monti S, Conson M, Pacelli R, Cella L. Normal tissue complication probability (NTCP) models for modern radiation therapy. Semin Oncol. 2019 Jun;46(3):210–8.
72. Blanchard P, Wong AJ, Gunn GB, Garden AS, Mohamed ASR, Rosenthal DI, et al. Toward a model-based patient selection strategy for proton therapy: External validation of photon-derived normal tissue complication probability models in a head and neck proton therapy cohort. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2016 Dec;121(3):381–6.
73. Stieb S, Lee A, van Dijk LV, Frank S, Fuller CD, Blanchard P. NTCP Modeling of Late Effects for Head and Neck Cancer: A Systematic Review. Int J Part Ther. 2021;8(1):95–107.
74. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013 Feb 5;10(2):e1001381.

Appendix

A1) Arriving at the outcome of interest and defining the research question

Phase I

A comprehensive search was conducted on MEDLINE and was split in 2 phases. In phase I, different search strategies were applied using Boolean and proximity operators to look for relevant post-radiogenic diseases. Various medical subject headings (MeSH)-terms were employed, namely: radiation neck cancer, radiotherapy neck cancer, post-radiogenic disease neck, post-radiogenic disease pharynx, nasopharyngeal carcinoma radiotherapy, oral cavity radiotherapy. Grounded on the results of the search, criteria were adjusted, and new queries were started. In parallel, the data set of the training cohort was scrutinized given that it was mandatory to find a potential outcome with sufficient amount of data (number of patients responses in total) and enough outcomes (patients suffering from the disease of interested) 6 months after the end of the RT. The index date was deliberately selected to have a clear demarcation between chronic and acute processes, with the latter having their onset within the first 3 months. Chronic disease was chosen because it was considered to be of greater clinical importance than an acute disease, which often possesses only short-term effects.

The MEDLINE search yielded n=20 different complications after RT for the treatment of HNCs. A closer inspection of these 20 post-radiogenic sequelae disclosed 3 disease complexes that frequently crop up after RT, namely: (I) xerostomia and/or dry mouth, (II) dysphagia, swallowing disorder, aspiration, and/or dependence on a feeding tube, and (III) hypothyroidism. Given that already a couple of NTCP-models had been set up and updated for these 3 complexes, they were omitted from the search list of potential outcomes of interest. Consequently, the next step focused on the 17 remaining complications after RT and grounded on data availability in the training data set, the search was narrowed to 6 diseases. Following, the clinical relevance of these 6 sequelae was discussed and complications without a distinct clinical relevance were sorted out. Ultimately, 2 diseases were shortlisted, namely: (I) trismus and/or reduced mouth opening and (II) laryngeal edema. Due to the lack of clear reference values of mouth opening, laryngeal edema was eventually selected as target disease and outcome for detailed analysis.

Phase II

After having chosen post-radiogenic laryngeal edema as disease of interest, the research question was clarified and specified in phase II. For this purpose, another MEDLINE search was executed, aiming to identify already existing NTCP-models as well as relevant factors for the prediction of post-radiogenic edema. This approach of first finding existing models, then validating them externally and eventually updating the models was in line with the recommendations of the Prognosis Research Strategy (PROGRESS) (2,74). Different Boolean and proximity operators were applied in the search and the following MeSH terms were used: laryngeal edema radiation, post-radiogenic laryngeal edema, and laryngeal edema radiotherapy. The publications detected were first looked through based on their headings and date of publication, and then initially scanned based on their abstracts. Eventually, the publications considered of being relevant to the research question were fully evaluated. Figure 3 gives an overview of the search tract.

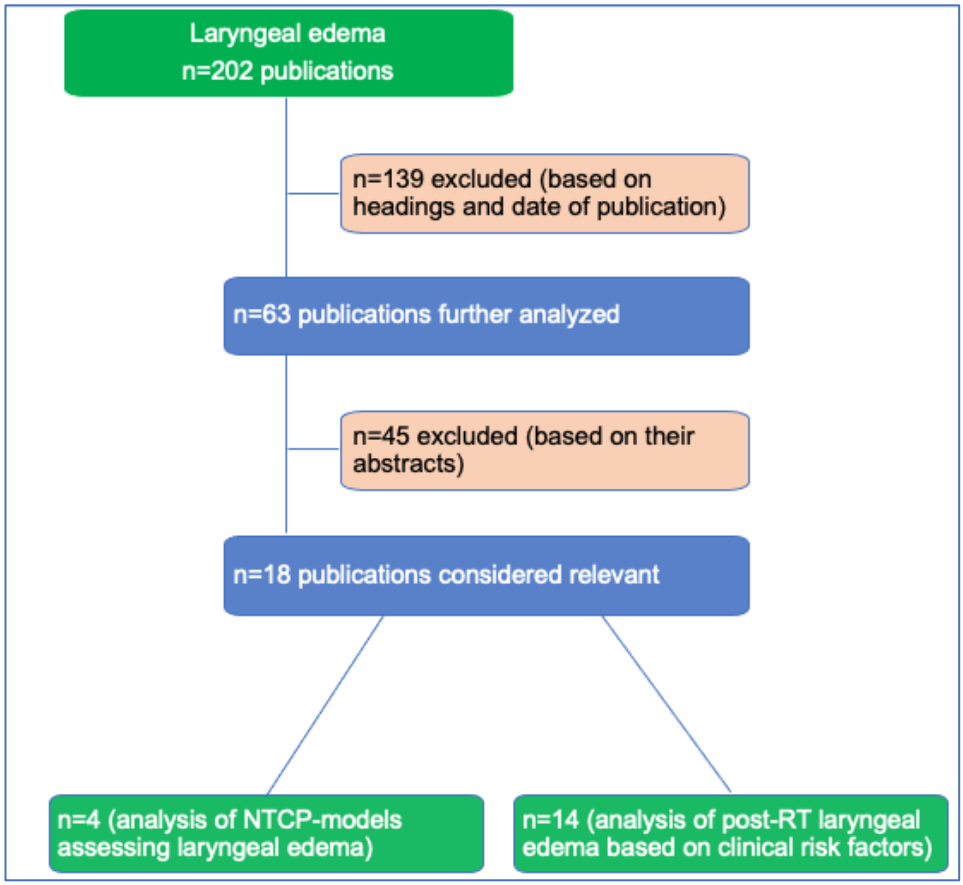

Figure 3. Chart of the search strategy applied to derive information for NTCP-model development in laryngeal edema.

Literature search yielded 4 studies that incorporated NTCP-models for the prediction of post-radiogenic laryngeal edema (8,33–35). Rancati et al. had been the first to introduce 2 NTCP-models dealing with this complication (13,16,32,33) and these models have been applied without any update in the other publications (33–35). Thus, until today only the 2 models published by Rancati et al. were identified in the literature (8,62).

Information on clinical risk factors for the occurrence of post-radiogenic laryngeal edema was provided in the remaining 14 of the selected 18 publications and formed the basis for the choice of the candidate predictors.

A2) Approximation of the EUD

In the present data set, the irradiation data of the larynx was separated into a supraglottic area and a glottic area, which were given as DVHs. This data was taken to approximate the EUD for each patient, inferring the d_i and v_i from the DVHs. The following irradiation data was available for each area: Volume, DMEAN (= mean dose), DMIN (= minimum dose), DMAX (= maximum dose), V01, V05, V10, V15, V20, V25. V30, V35, V40, V45, V50, V55, V60, V65, V70, V75, V80, V85. The combination of V and a number reflects the volume, which has received at least the respective dose, e.g., V50=23% of the glottic area means, that 23% of the glottic volume has been given a dose of at least 50 Gy. Considering this information, the volume parameters (v_i in the Rancati model) were approximated based on the following assumptions (example data):

- V20=42%
- V10=50%
- V5=100%

It can be inferred that a volume of 100-50%=50% received a dose between 5 and 10 Gy (using V5 and V10); thus, D=7.5 and v=0.5 are the first "pair"; a volume of 50-42%=8% were irradiated with a dose between 10 and 20 Gy (using V10 and V20); therefore, D=15 and v=0.08 are the next "pair". The same procedure can be carried out for all existing dose values in a specific area, so that in the end all pairs of v_i and d_i necessary for calculating the EUD become available. The present dose data did not include a patient who received a single dose of 85 Gy (V85). Thus, the current approximation is limited between 0 Gy and 85 Gy, thereby reducing the risk of inaccuracy in calculating the upper limit.

The given values for the constants from the best-fit model from Rancati et al. (8) were taken over, meaning n=1.41, k=7.2 and D_{50}=46.7 were used in the EUD calculation.

A3) Univariable analysis for each candidate variable

At the beginning of the model development process, a univariable analysis of the candidate predictors was performed, the results are summarized in Table 14 below. For non-numeric variables ranges of the estimate, standard error, and p-values are given to reflect the different levels of the predictor.

Variable	**Estimate**	**Standard error**	**AIC**	**p-value**
EUD	0.44	0.17	207.6	< 0.01
BASELINE EDEMA	-14.03 - 6.08	1073.11 - 1678.46	211.13	1
SMOKING	1.6	1.03	210.43	0.12
HPV	13.07 - 14.05	979.61	214.34	0.99
SEQUENCE	-14.02 - 2.16	0.58 - 932.48	200.07	< 0.01-0.99
LOCATION	-15.52 - 1.24	0.77 - 1190.87	209.05	0.12-0.99
T-STAGE	-9.75 - 10.25	614.27 - 1300.69	200.92	0.99-1
AGE	0.04	0.02	209.31	0.03
HISTOLOGY	-16.06	1190.87	209.9	0.99
SUPRAGLOTTIC DMAX	0.3	0.09	194.94	< 0.01
GLOTTIC DMAX	0.11	0.03	190.17	< 0.01
SUPRAGLOTTIC DMEAN	0.14	0.03	181.19	< 0.01
GLOTTIC DMEAN	0.09	0.2	187.48	< 0.01
TECHNIQUE	-13.68 - 0.35	1.05 - 848.37	217.05	0.74-0.99

Table 14. Summary of the univariable analysis of the 14 candidate predictors.
AIC= Akaike Information Criterion

The variables BASELINE EDEMA, HPV, T-STAGE, HISTOLOGY, and TECHNIQUE were subsequently excluded from further analysis as none of their factor levels revealed a p-value <0.2.