Acta Oncologica, 2015; 54: 1301–1309

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

External validation of a normal tissue complication probability model for radiation-induced hypothyroidism in an independent cohort

MARIANNE F. RØNJOM[1,2], CARSTEN BRINK[2,3], SØREN M. BENTZEN[4,5], LASZLO HEGEDÜS[2,6], JENS OVERGAARD[7], JØRGEN B. B. PETERSEN[8], HANNE PRIMDAHL[9] & JØRGEN JOHANSEN[1,2]

[1]*Department of Oncology, Odense University Hospital, Odense, Denmark,* [2]*Institute of Clinical Research, University of Southern Denmark, Odense, Denmark,* [3]*Laboratory of Radiation Physics, Odense University Hospital, Odense, Denmark,* [4]*Division of Biostatistics and Bioinformatics, University of Maryland Greenebaum Cancer Center, Baltimore, USA,* [5]*Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, USA,* [6]*Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark,* [7]*Department of Experimental Clinical Oncology, Aarhus University Hospital, Aarhus, Denmark,* [8]*Department of Medical Physics, Aarhus University Hospital, Aarhus, Denmark and* [9]*Department of Oncology, Aarhus University Hospital, Aarhus, Denmark*

ABSTRACT

Background. A normal tissue complication probability (NTCP) model for radiation-induced hypothyroidism (RIHT) was previously derived in patients with squamous cell carcinoma of the head and neck (HNSCC) discerning thyroid volume ($V_{thyroid}$), mean thyroid dose (D_{mean}), and latency as predictive factors. The purpose of this study was to test the performance of this model in an independent cohort of patients receiving primary radiotherapy (RT) for HNSCC.

Material and methods. A validation cohort of 198 patients with HNSCC was included after plasma thyrotropin (TSH) assessment. RIHT was defined as TSH > 4.0 mU/l from blood samples obtained during follow-up. A new mixture NTCP model was developed from the validation cohort after multivariable analysis. Due to only one follow-up TSH assessment in the validation cohort, the time factor derived from the original cohort was fixed in a mixture model and applied for the NTCP validation. Association between model predictions of the initial model and observed clinical outcome in the validation cohort was investigated by applying the previous model ($V_{thyroid}$, D_{mean} and time) on the new cohort and comparing it to the clinical outcome.

Results. Both D_{mean} and $V_{thyroid}$ were confirmed as significant risk factors for RIHT in the validation cohort, odds ratio (OR) 1.19 (1.1–1.37) and OR 0.75 (0.57–0.9), respectively. A small difference in overall probability of RIHT was observed between the cohorts, further analysis indicated this to be related to less frequent blood tests in the validation cohort relative to the original cohort. However, Pearson's correlation coefficients between model and clinical outcome were high: $r = 0.97$ estimated by the original model versus the original cohort, and $r = 0.97$ estimated by the original model versus the new cohort.

Conclusion. D_{mean} and $V_{thyroid}$ were significant predictors of RIHT in both cohorts. The original NTCP model demonstrated external validity owing to high Pearson's correlation coefficients between estimated and observed incidence rates of RIHT in the original as well as in the validation cohort. This model may facilitate clinically relevant estimations of RIHT after RT to the neck.

Radiation-induced hypothyroidism (RIHT) is a well known late effect after primary radiotherapy (RT) to the neck area [1–5]. An association between hypothyroidism and morbidity and mortality has been established [6–11] for which reason protective measures should be undertaken during treatment

Correspondence: J. Johansen, Department of Oncology, Odense University Hospital, Denmark. E-mail: jorgen.johansen@rsyd.dk

(Received 15 May 2015; accepted 12 June 2015)

ISSN 0284-186X print/ISSN 1651-226X online © 2015 Informa Healthcare
DOI: 10.3109/0284186X.2015.1064160

planning for RT of the head and neck to prevent damage to the thyroid gland. Due to large heterogeneities in previous retrospective studies, no clear picture has emerged to describe specific levels of glandular dysfunction in relation to radiation treatment doses. This has been accomplished in recent studies of patients with squamous cell carcinoma of the head and neck (HNSCC) receiving RT as the primary treatment [1–3]. We developed a normal tissue complication probability NTCP model for RIHT [1] including both thyroid volume ($V_{thyroid}$) and mean radiation dose to the thyroid gland (D_{mean}) and estimated a five-year risk of developing RIHT of 26% for the whole study population with RIHT mainly evolving within three years after RT. Two additional studies have proposed other NTCP models for development of RIHT, all consistently identifying radiation dose [2,3] and $V_{thyroid}$ [2] as significantly associated with RIHT. From such dose response-modeling, recommendations with respect to dose/volume constraints in dose planning have been proposed [1].

However, it is important to evaluate the external validity of such NTCP models before eventual implementation in routine clinical practice. As stated in The Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) report [12], too few external validations are actually performed. Therefore, we aimed to test the performance of our original NTCP model in an independent cohort of patients with HNSCC from another institution.

Material and methods

The original cohort (cohort 1) treated at Odense University Hospital (OUH), Denmark, has been described previously [1]. A similar sized validation cohort (cohort 2) of patients with HNSCC was obtained from the Department of Oncology at Aarhus University Hospital, Denmark. Patients were included after an assessment of plasma thyrotropin (TSH) at routine follow-up in the outpatient clinic from June to December 2012 until a cohort about the same size as cohort 1 was obtained (Supplementary Figure 1, available online at: http://informahealthcare.com/doi/abs/10.3109/0284186X.2015.1064160). At the department in Aarhus, patients are seen in the outpatient clinic at regular monthly intervals up to five years after primary radiotherapy.

The inclusion and exclusion criteria were identical in the two studies: invasive HNSCC of the oral cavity, oropharynx, hypopharynx or larynx, treated according to the Danish Head and Neck Cancer Group (DAHANCA) guidelines with external beam RT, with a planned total dose of 66–68 Gy in 33–34 fractions, 5–6 fractions/week, six fractions as BID one day weekly. In cohort 1, patients receiving 68 Gy with IMRT could be treated in 33 fractions to assure a dose per fraction of at least 1.5 Gy to elective sites.

Concomitant weekly platinum (cisplatin 40 mg/m^2 or carboplatin AUC 1.5) with or without the radiosensitizer nimorazole (1200 mg/m^2) was allowed. However, no primary surgery was allowed due to the increased risk of RIHT from this treatment modality [5]. Patients were excluded in case of a new primary HNSCC or retreatment of recurrent disease, except salvage treatment with neck dissection alone. Patients with nasopharyngeal cancer were excluded due to the possible impact of radiation on the pituitary gland and consequent development of central RIHT. A computed tomography (CT)-based three-dimensional (3D) treatment plan had to be available for dose calculation, excluding patients receiving an electron boost. Patients who had received previous RT to the neck area were excluded if the thyroid D_{mean} exceeded 2 Gy or the exact thyroid dose could not be calculated.

Endocrine assessment

Patients were defined as euthyroid if a TSH measurement was within the normal range (TSH 0.3–4.0 mIU/l [13]). In the present study, biochemical RIHT was defined as a measured TSH above > 4.0 mIU/l at any time after RT.

To avoid bias of TSH assessments, a standard procedure was applied. All blood samples were immediately centrifuged, pipetted into tubes, frozen and sent to the laboratory at OUH where they were analyzed. Identically to cohort 1 TSH analyses were done using Architect *i* system, Abbot. Baseline levels of TSH were obtained from frozen blood samples taken routinely before start of RT, or within the first two weeks of treatment. Blood samples were obtained from the blood bank at the Department of Experimental Clinical Oncology, Aarhus University Hospital (lithium-heparin plasma stored at -80 °C). Analyses were performed in March 2013. Baseline assessments in cohort 2 were obtained before RT in five patients, during the first week of RT in 150, and during the second week in the remaining 43 patients.

One baseline and one follow-up sample of TSH were obtained from cohort 2.

Treatment planning/radiotherapy

Treatment planning CT-scans were performed on a Philips Gemini TF 16 Big Bore PET-CT scanner or a Philips Brilliance BB16 CT scanner, both with a slice thickness of 3 mm. Treatment planning was

done using Eclipse (Varian Medical Systems, Inc., Palo Alto, CA, USA) version 8.0, 8.5, 8.6 or 10.0. Delineation of the thyroid gland and dose calculation in all plans was done using Eclipse version 10.0. All patients were treated in the period 2007–2012 and the majority of the treatment plans were done as IMRT and doses were calculated using tissue heterogeneity. Sparing the thyroid gland was not part of the optimization objectives in contrast to the adjacent larynx and swallowing structures. The treatment plans were exported as DICOM RT files to the DICOM Collaboration system [14] at Odense University Hospital and D_{mean} and $V_{thyroid}$ were extracted from this server.

Statistical analysis and NTCP modeling

The binary endpoint was presence of biochemical RIHT as defined above. Follow-up time was defined from the last day of RT to the date of blood draw. For comparison of the characteristics of patients included in cohorts 1 and 2, the Mann Whitney U-test or Fisher's exact test were used when appropriate. In our previous study [1], a mixture model was used for multivariable analyses and NTCP modeling. The mixture model was a logistic dose response model incorporating covariates affecting the risk of RIHT and a latent time distribution parameterized by a cumulative Weibull distribution from which actuarial risk estimates of hypothyroidism could be obtained.

The NTCP model previously developed from cohort 1 was:

$$NTCP = (1 + \exp(-S))^{-1}, \text{ where:}$$

$$S = k_{const} + k_{Dose} * D_{mean} + k_{Vol} * V_{thyroid}$$

More than half of the patients in cohort 1 had multiple (2–11) TSH measurements during follow-up. The serial blood samples were used to assess the time to development of RIHT and from this, incorporating the time factor in the analysis. Model optimization of the data in cohort 2 with just one TSH assessment per patient was unattainable when the parameters related to the latent time distribution were included in the optimization due to very large uncertainties of the fitted values. Therefore, the latent time distribution parameters were fixed at the estimates from cohort 1 when analyzing the RIHT data from cohort 2.

In addition to the validation of the original model applied to cohort 2, multivariable mixture model analysis was performed in cohort 2 including volume ($V_{thyroid}$) and mean dose (D_{mean}) as in the previous study, and with the fixed latent-time distribution. The following covariates: baseline TSH, age, site (dichotomized), stage (dichotomized), neck dissection after RT, concomitant ChRT, and sex were added stepwise to the cohort 2 model to test for statistical significance. 95% CIs were derived by the profile likelihood method, and p-values for significance were derived from the likelihood ratio (LR) test. A new mixture NTCP model, NTCP2, was derived from the multivariable analysis.

Validation of the NTCP model was performed using calibration plots showing the clinically observed outcome in groups of patients binned according to the predicted outcome. This was achieved by estimating the individual patient's predicted risk of RIHT at the time of being diagnosed with RIHT, or at the time of the last blood draw. The predicted values were then ranked in ascending order and used to form nine equally sized groups with increasing predicted risk of RIHT. The observed probability of RITH within each group was plotted as a function of mean estimated risk for each group. Error bars of observed risks were calculated assuming a binomial distribution. Correlation between the predicted and observed values was quantified by Pearson's correlation coefficient.

An explorative approach was executed to investigate the influence of the number of TSH samples on the predictions. This could either be done by randomly selecting a single blood test from each patient in cohort 1 to create a new NTCP model and subsequently evaluate whether this model was closer to the NTCP model derived from cohort 2 than the initial model from cohort 1. However, a single sample from cohort 1 might induce biased estimates if "extreme" data values had been sampled by coincidence. To address this issue, subsamples from cohort 1 (containing only one blood test per patient) were sampled not just once, but 10 000 times. Each of these 10 000 subsets of cohort 1 was used to make 10 000 new NTCP models, referred to as cohort 1S (Monte Carlo sampled cohort). Cohort 1S thus generated a distribution of values, 10 000 values, for each NTCP model parameter. The Cohort 1S values are therefore reported as the median value of the parameter distribution and the 95% CIs were obtained as the central region (2.5% below and above) that includes 95% of parameter values from the Monte Carlo simulation.

Ethics

The study was approved by the Regional Ethics Committees for Southern Denmark (S-20110027) and the Danish Data Protection Agency (J.nr. 2012-41-0899).

Results

A total of 198 patients were included in cohort 2. There were no statistically significant differences between the two cohorts as regards sex, age, primary tumor site, stage or $V_{thyroid}$. D_{mean} was significantly lower in cohort 2, median 33.5 Gy compared with median 39.2 Gy in cohort 1, $p = 0.018$ (Table I). Baseline TSH was also significantly lower in cohort 2, $p < 0.001$.

Median time to TSH assessment in cohort 2 was 22.4 months (range 0.5–62.6) after RT. Nineteen of 198 (9.6%) patients in cohort 2 had biochemical RIHT on the follow-up test, 12 patients had subclinical HT [TSH above the normal range and free (F)-T4 within the normal range of 9.9–17.7 pmol/l] and seven patients had developed overt HT (TSH above and F-T4 below the normal range).

D_{emean} and $V_{thyroid}$ were consistently found as highly significant risk factors for development of RIHT from the multivariable analyses in all cohorts (Table II). In cohort 2, baseline TSH was also a significant risk factor. The model was also analyzed with the addition of baseline TSH, $V_{thyroid}$ and D_{mean} still retained significance (data not shown). None of the other factors were significant in the multivariable analyses.

The model parameters for the original NTCP model (NTCP1) and the NTCP model from the validation cohort (NTCP2) with fixed latent time distribution are shown in Table III. Figure 1 demonstrates the dose response curves for the two NTCP models for selected thyroid volumes.

Calibration plots of predicted results in cohorts 1 and 2 are shown in Figure 2A and B. The calibration plots demonstrated very high Pearson's correlation coefficients between predicted and observed risk (cohort 1 vs. NTCP1, $r = 0.97$; cohort 2 vs. NTCP2, $r = 0.96$). This also applied to the calibration of the

Table I. Patient characteristics of the original cohort (cohort 1) and validation cohort (cohort 2).

Variable	Cohort 1 number (%)	Cohort 2 number (%)	p-Value
Sex			
Male	157 (77.3)	155 (78.3)	NS[b]
Female	46 (22.7)	43 (21.7)	
Age (years)			
Median (range)	61.1 (41.5–86.8)	60.2 (31.4–85.0)	NS[a]
Tumor site			
Oral cavity	7 (3.5)	6 (3.0)	NS[b]
Oropharynx	102 (50.3)	100 (50.5)	
Hypopharynx	11 (5.4)	10 (5.1)	
Larynx	83 (40.9)	82 (41.4)	
Stage			
I+II	84 (41.3)	66 (34.2)	NS[b]
III+IV	119 (58.7)	127 (65.8)	
Dose (Gy) to primary site			
60		3 (1.5)	0.021[b]
62		1 (0.5)	
66	158 (77.8)	134 (67.7)	
68	45 (22.2)	60 (30.8)	
Number of fractions			
30		2 (1)	<0.001[b]
33	191 (94.1)	143 (71.8)	
34	12 (5.9)	55 (27.8)	
Systemic treatment			
None	159 (77.8)	120 (60.6)	<0.001[b]
Chemotherapy	44 (22.2)	78 (39.4)	
Mean thyroid dose (Gy)			
Median (range)	39.2 (1.2–68.0)	33.5 (3.8–64.8)	0.018[a]
Thyroid volume (cm^3)			
Median (range)	17.4 (5.4–51.5)	17.3 (6.4–85)	NS[a]
Time to (last) follow-up (months)			
Median (range)	25.1 (1.8–116.3)	22.4 (0.5–62.6)	
Baseline TSH (mU/l)			
Median (range)	1.24 (0.37–3.7)	0.93 (0.3–3.43)	<0.001[a]
Neck dissection after RT			
No	172 (84.7)	137 (69.2)	
Yes	31 (15.3)	61 (30.8)	<0.001[b]

[a]analyzed with the Mann-Whitney U-test; [b] analyzed with Fisher's exact test.

Table II. Results of multivariable analyses of RIHT in the original cohort 1, validation cohort 2 and the sampled cohort 1S.

	Cohort 1			Cohort 2			Cohort 1S		
	OR	95% CI	p-Value	OR	95% CI	p-Value	OR	95% CI	p-Value
Mean thyroid dose (Gy)									
Continuous	1.12	1.07–1.20	<0.001	1.19	1.10–1.37	<0.001	1.20	1.11–1.45	<0.001
Thyroid volume (cm^3)									
Continuous	0.75	0.64–0.85	<0.001	0.75	0.57–0.90	<0.001	0.74	0.54–0.86	<0.001
Baseline TSH									
Continuous	1.84	0.79–5.09	NS	4.69	1.45–21.68	0.01	2.42	0.70–15.41	NS
Age (years)									
Continuous	0.99	0.93–1.06	NS	0.97	0.84–1.07	NS	0.98	0.88–1.07	NS
Tumor site[a]									
1 + 2	0.95	0.27–3.18	NS	1.17	0.26–5.20	NS	1.53	0.2–10.96	NS
3 + 4									
Stage									
I + II	0.48	0.10–1.83	NS	1.84	0.23–17.97	NS	0.22	0.00–2.44	NS
III + IV									
Neck dissection after RT									
None	1.06	0.23–5.01	NS	0.34	0.06–1.73	NS	0.86	0.08–16.74	NS
Neck diss.									
Systemic treatment									
None	0.51	0.12–1.94	NS	2.17	0.48–13.23	NS	0.46	0.04–5.17	NS
Chemotherapy									
Sex									
Male	1.26	0.29–5.42	NS	0.94	0.14–6.17	NS	0.62	0.02–8.77	NS
Female									

[a]1 = oral cavity, 2 = oropharynx, 3 = hypopharynx, 4 = larynx.

original model on the new, independent cohort (cohort 2 vs. NTCP1, r = 0.97) (Figure 2C).

We found higher risk estimates of RIHT by NTCP1 (multiple blood samples) than actually observed in cohort 2 (Figure 2C) (slope of calibration plot less than unity). This was also observed when NTCP1 and NTCP2 were compared (Figure 3A).

In the additional analyses of a supplementary sampling of the original patient cohort, cohort 1S, imitating a single blood test sampling, demonstrated that only D_{mean} and $V_{thyroid}$ were significant risk factors for RIHT in the sample, as in cohort 1 (Table II) and the corresponding parameters of the NTCP model (NTCP1S) are shown in Table III. Thus, comparison of NTCP1S and NTCP2 showed almost identical risk estimates between the two one-sample cohorts (Figure 3B).

Discussion

RIHT is a frequent late effect after RT with reported incidences of biochemical RIHT ranging from around 20% to over 50% [1–4]. The incidences are considerably increased compared to the background population [15,16]. Bearing in mind the possible long-term consequences of hypothyroidism, such as increased morbidity and mortality [6–10], as well as decreased quality of life [11], there is a need to establish the dose response relationship for RIHT as a rational way to decide on dose constraints for treatment planning. While we previously have given recommendations with respect to dose-volume constraints for the thyroid gland in head and neck cancer patients, we also found large confidence intervals in our study population [1]. Therefore, larger cohorts of patients are needed to improve estimates for clinical application, and dose response relationships – as described by NTCP models – should attain a robustness that allows generalization between different patient populations.

Table III. Model parameters from the original NTCP model (NTCP1), the NTCP model from the validation cohort (NTCP2) and the Monte Carlo sampled model (NTCP1S).

Model	k_{const}	k_{Dose}	k_{Vol}
NTCP1	−1.24	0.12	−0.29
NTCP2	−4.71	0.17	−0.28
NTCP1S	−4.92	0.18	−0.30

There are considerable uncertainties associated with NTCP models. We have tried to minimize the impact of random effects on estimates of RIHT by including a fairly large number of patients in two independent cohorts of patients with HNSCC. Furthermore, patients in both cohorts received a relatively standardized treatment according to the DAHANCA guidelines [17], and no surgery was allowed to reduce the impact of other well known risk factors for hypothyroidism.

The two cohorts studied here were considered as representative for the population of patients with

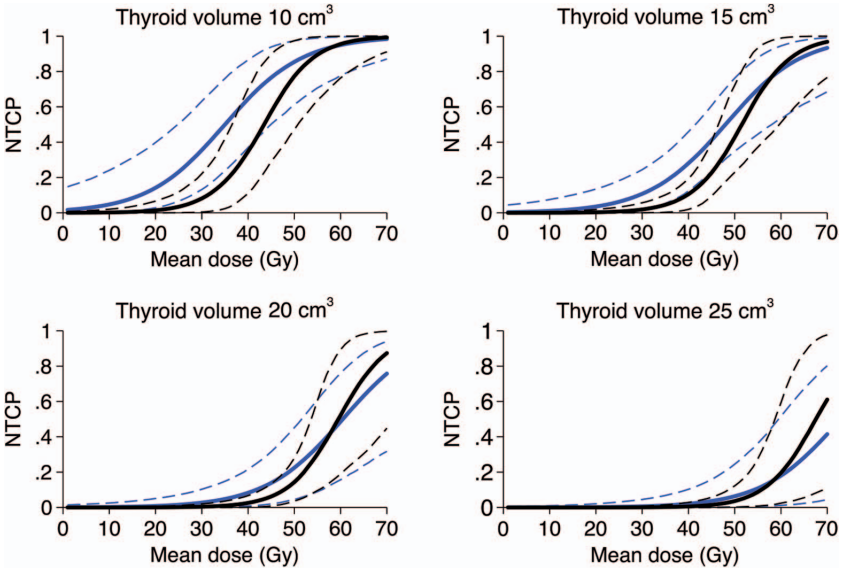

Figure 1. Dose response curves for the original NTCP model (NTCP1) in blue and the model from the validation cohort (NTCP2) in black for selected thyroid volumes.

HNSCC treated with primary RT at the two institutions, however, the median D_{mean} was significantly lower in cohort 2 than in cohort 1. This may be ascribed to a dose planning strategy used in the department in Aarhus to lower the dose to the larynx and swallowing structures (to reduce the risk of radiation-induced dysphagia), subsequently leading to a lower dose to the adjacent thyroid gland.

To minimize another potential analytical bias between the cohorts in estimating our endpoint, all blood samples were analyzed in the same laboratory and the use of frozen blood samples in cohort 2 was considered appropriate according to other investigations [18]. However, a significant difference in baseline TSH was found. This might at least partly be explained by the time of blood sampling during the first two weeks of treatment in most of the patients in cohort 2 in contrast to only eight in cohort 1, as thyroiditis with subsequent reduced TSH levels is a recognized acute reaction from RT to the neck [19].

By taking latency into account in NTCP modeling, the time of occurrence of an event is estimated as well as time censoring [1,20]. Bentzen et al. [21] demonstrated that it was possible to estimate latency from a single follow-up clinical study. However, in the current study, the available data regarding time in relation to only one TSH assessment did not allow us to repeat such an estimate of the latency function, as the uncertainties were too large for a meaningful comparison of the two cohorts. Therefore, the latency function of cohort 1 was applied on cohort 2 and,

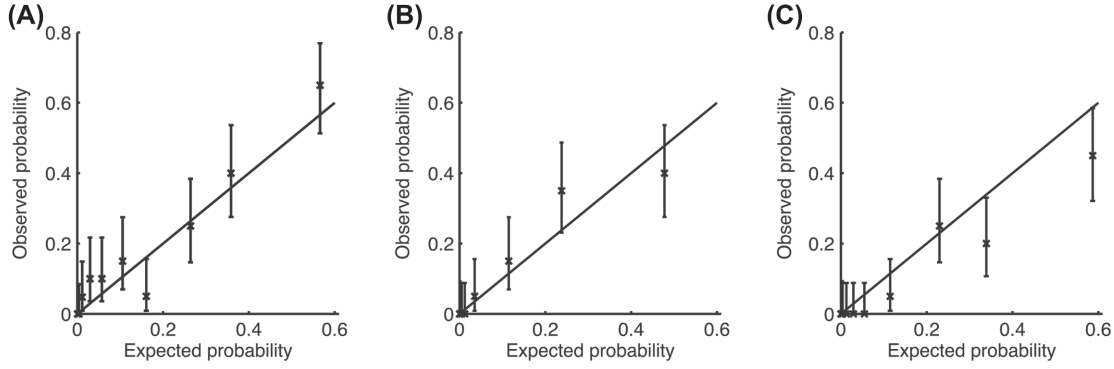

Figure 2. Calibration plots (A) demonstrates the expected probability from the original NTCP1 versus the observed probability in the original cohort 1, (B) expected probability from NTCP2 versus observed probability in validation cohort 2, C) Demonstrates the external validation of the original NTCP model, expected probability from NTCP1 (estimated on cohort 2) versus observed probability in validation cohort 2.

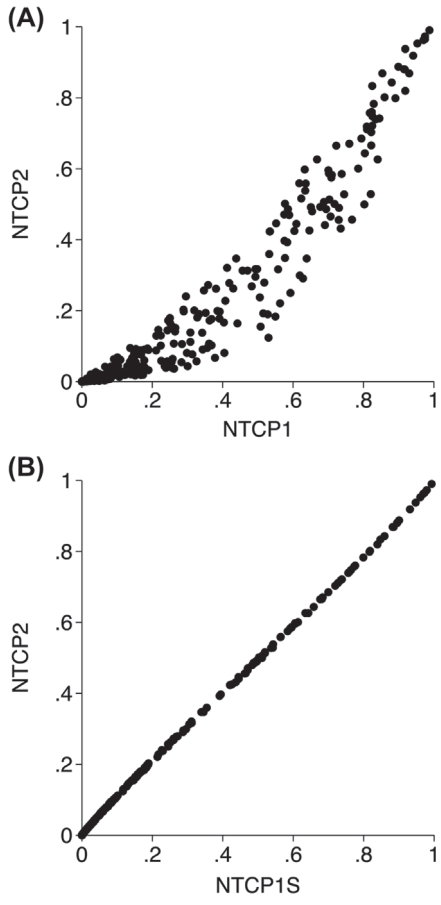

Figure 3. (A) The risk of developing radiation-induced HT, estimated from the original NTCP1 (multiple blood samples) and NTCP2 from the validation cohort (single blood samples). B) The risk of HT, estimated from the Monte Carlo sampled NTCP1S and NTCP2 (both one blood sample). Estimations were done on patients from both cohorts.

the validity of the latency function was thus not validated in the current study.

In the QUANTEC report, three aspects of validation of NTCP models were distinguished: face validity, internal validity and external validity [12]. Face validity relates to whether the findings of a study are reasonable. Pretreatment $V_{thyroid}$ and D_{mean} were found to be significant risk factors in cohort 2 as well as in cohort 1 and cohort 1S, which is in accordance with other studies [2,3,22]. Sex has been found to be a significant risk factor in earlier studies [5], which may reflect that women have smaller thyroid glands [23]. In the original study of cohort 1, volume was still significant in subgroup analyses of men and women separately, while sex as such was not an independent factor [1]. Baseline TSH was a significant risk factor in multivariable analyses in cohort 2, but not in cohort 1. Baseline TSH was lower in cohort 2, which could be ascribed to blood sampling during the first two weeks of treatment, and was not included in the NTCP2 model, since baseline TSH was not associated with RIHT in cohort 1. From the results of our multivariable analyses, in relation to other studies, we can conclude that the findings in the two cohorts and the derived NTCP models have face validity.

The internal model validity, which is the ability of the model to predict the clinical outcome of the data used for modeling, is shown in Figure 2A and B. High Pearson's correlation values were found demonstrating that the models predict the relative risk probabilities with high precision. Furthermore, the data points in Figure 2A and B scatter around the line of identity, which shows that also the absolute predicted risk values are in line with the clinically measured values.

The aim of this study was to carry out an external validation of our mixture NTCP model (NTCP1) that is the ability of a model to predict clinical outcome in a separate cohort from the one used for the initial modeling. The calibration plot in Figure 2C shows the external validity of this model. The correlation coefficient between predicted values from NTCP1 and those observed in cohort 2 is very high, and the calibration plot demonstrates that NTCP1 has external validity in terms of relative risk probabilities.

If the absolute NTCP is considered (Figure 2C), there is a tendency that NTCP1 generates higher risk estimates of RIHT than is observed in cohort 2 (also apparent when comparing the dose response curves from NTCP1 and NTCP2 in Figure 1). The small difference in predicted absolute values is not necessarily related to biological differences between the two cohorts. This is observed in the simulation of cohort 1S that is a simulation of a likely outcome if only one blood sample per patient had been present during the initial NTCP modeling. The NTCP1S is almost identical to the NTCP2 model (Table III). This shows that the small difference in prediction of the absolute values of the original model of cohort 2 might be related to the use of only one blood sample resulting in fewer events since the measurement of just a single TSH above 4 is defined as an event. Diagnosing RIHT from a single TSH measurement above 4.0 mIU/l at any time, considering the biological and analytical variation in TSH assessments [24], may lead to an overestimation of RIHT. However, this endpoint is commonly used in studies of RIHT [2,3,5,22].

The independent factors in the models of RIHT are volume of the thyroid gland and the mean dose to the thyroid. These are obviously dependent of the delineation accuracy of the thyroid gland. We investigated the impact of delineation of the gland by different observers [25] and found that the variation in predicted risk of radiation-induced HT was rather

small for the general study population and concluded that the proposed NTCP model is relatively insensitive to delineation uncertainties of the thyroid gland, which is important for the use of the model in other institutions.

In conclusion, in this validation cohort, we confirmed decreasing $V_{thyroid}$ and increasing mean dose to the thyroid gland as significant risk factors for RIHT after RT for head and neck cancer. When comparing the predicted risk of RIHT by the original NTCP model in the new cohort, the predictions correlated strongly with clinical observations, demonstrating the external validity of the model. A minor difference in overall predicted risk values is possibly related to different number of blood samples in the two cohorts. Considering the possible long-term consequences of hypothyroidism it is important that the thyroid gland is regarded as an organ at risk in treatment planning of RT. The validation of our NTCP model has shown that the model reliably describes the dose response relationship for RIHT.

Acknowledgments

This study was supported by CIRRO – The Lundbeck Foundation Center for Interventional Research in Radiation Oncology and the Danish Cancer Society, The A.P. Møller Foundation for the Advancement of Medical Science, "Frimondt-Heineke Fonden", "Fabrikant Einar Willumsens Mindelegat", "Lily Benthine Lunds Fond af 1.6.1978" and" Torben og Alice Frimots Fond". Laszlo Hegedüs is supported by an unrestricted grant from the Novo Nordisk Foundation.

Declaration of interest: The authors report no conflict of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Rønjom MF, Brink C, Bentzen SM, Hegedüs L, Overgaard J, Johansen J. Hypothyroidism after primary radiotherapy for head and neck squamous cell carcinoma: Normal tissue complication probability modeling with latent time correction. Radiother Oncol 2013;109:317–22.

[2] Boomsma MJ, Bijl HP, Christianen ME, Beetz I, Chouvalova O, Steenbakkers RJ, et al. A prospective cohort study on radiation-induced hypothyroidism: Development of an NTCP model. Int J Radiat Oncol Biol Phys 2012;84:e351–6.

[3] Bakhshandeh M, Hashemi B, Mahdavi SR, Nikoofar A, Vasheghani M, Kazemnejad A. Normal tissue complication probability modeling of radiation-induced hypothyroidism after head-and-neck radiation therapy. Int J Radiat Oncol Biol Phys 2013;85:514–21.

[4] Boomsma MJ, Bijl HP, Langendijk JA. Radiation-induced hypothyroidism in head and neck cancer patients: A systematic review. Radiother Oncol 2011;99:1–5.

[5] Vogelius IR, Bentzen SM, Maraldo MV, Petersen PM, Specht L. Risk factors for radiation-induced hypothyroidism: A literature-based meta-analysis. Cancer 2011;117:5250–60.

[6] Thvilum M, Brandt F, Almind D, Christensen K, Brix TH, Hegedüs L. Type and extent of somatic morbidity before and after the diagnosis of hypothyroidism. A nationwide register study. PLoS One 2013;8:e75789.

[7] Thvilum M, Brandt F, Almind D, Christensen K, Brix TH, Hegedüs L. Increased psychiatric morbidity before and after the diagnosis of hypothyroidism: A nationwide register study. Thyroid 2014;24:802–8.

[8] Rodondi N, den Elzen WP, Bauer DC, Cappola AR, Razvi S, Walsh JP, et al. Subclinical hypothyroidism and the risk of coronary heart disease and mortality. JAMA 2010;304:1365–74.

[9] McQuade C, Skugor M, Brennan DM, Hoar B, Stevenson C, Hoogwerf BJ. Hypothyroidism and moderate subclinical hypothyroidism are associated with increased all-cause mortality independent of coronary heart disease risk factors: A PreCIS database study. Thyroid 2011;21:837–43.

[10] Thvilum M, Brandt F, Almind D, Christensen K, Hegedüs L, Brix TH. Excess mortality in patients diagnosed with hypothyroidism: A nationwide cohort study of singletons and twins. J Clin Endocrinol Metab 2013;98:1069–75.

[11] Watt T, Cramon P, Hegedus L, Bjorner JB, Bonnema SJ, Rasmussen AK, et al. The thyroid-related quality of life measure ThyPRO has good responsiveness and ability to detect relevant treatment effects. J Clin Endocrinol Metab 2014;99:3708–17.

[12] Bentzen SM, Constine LS, Deasy JO, Eisbruch A, Jackson A, Marks LB, et al. Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): An introduction to the scientific issues. Int J Radiat Oncol Biol Phys 2010;76:S3–9.

[13] Jensen E, Hyltoft Petersen P, Blaabjerg O, Hansen PS, Brix TH, Kyvik KO, et al. Establishment of a serum thyroid stimulating hormone (TSH) reference interval in healthy adults. The importance of environmental factors, including thyroid antibodies. Clin Chem Lab Med 2004;42:824–32.

[14] Westberg J, Krogh S, Brink C, Vogelius IR. A DICOM based radiotherapy plan database for research collaboration and reporting. J Phys Conf Series 2014;489:012100.

[15] Carle A, Laurberg P, Pedersen IB, Knudsen N, Perrild H, Ovesen L, et al. Epidemiology of subtypes of hypothyroidism in Denmark. Eur J Endocrinol 2006;154:21–8.

[16] Knudsen N, Bulow I, Jorgensen T, Laurberg P, Ovesen L, Perrild H. Comparative study of thyroid function and types of thyroid dysfunction in two areas in Denmark with slightly different iodine status. Eur J Endocrinol 2000;143:485–91.

[17] DAHANCA. Danish Head and Neck Cancer Group (DAHANCA) guidelines 2013 [updated 2013; cited 2014 Oct 28]. Available from: http://www.dahanca.dk/guidelines.

[18] Mannisto T, Surcel HM, Bloigu A, Ruokonen A, Hartikainen AL, Jarvelin MR, et al. The effect of freezing, thawing, and short- and long-term storage on serum thyrotropin, thyroid hormones, and thyroid autoantibodies: Implications for analyzing samples stored in serum banks. Clin Chem 2007;53:1986–7.

[19] Bakhshandeh M, Hashemi B, Mahdavi SR, Nikoofar A, Edraki HR, Kazemnejad A. Evaluation of thyroid disorders during head-and-neck radiotherapy by using functional analysis and ultrasonography. Int J Radiat Oncol Biol Phys 2012;83:198–203.

[20] Bentzen SM, Thames HD, Travis EL, Ang KK, Van der Schueren E, Dewit L, et al. Direct estimation of latent time for radiation injury in late-responding normal tissues: gut, lung, and spinal cord. Int J Radiat Biol 1989;55:27–43.

[21] Bentzen SM, Thames HD, Overgaard M. Latent-time estimation for late cutaneous and subcutaneous radiation reactions in a single-follow-up clinical study. Radiother Oncol 1989;15:267–74.
[22] Diaz R, Jaboin JJ, Morales-Paliza M, Koehler E, Phillips JG, Stinson S, et al. Hypothyroidism as a consequence of intensity-modulated radiotherapy with concurrent taxane-based chemotherapy for locally advanced head-and-neck cancer. Int J Radiat Oncol Biol Phys 2010;77:468–76.
[23] Alterio D, Jereczek-Fossa BA, Franchi B, D'Onofrio A, Piazzi V, Rondi E, et al. Thyroid disorders in patients treated with radiotherapy for head-and-neck cancer: A retrospective analysis of seventy-three patients. Int J Radiat Oncol Biol Phys 2007;67:144–50.
[24] Andersen S, Bruun NH, Pedersen KM, Laurberg P. Biologic variation is important for interpretation of thyroid function tests. Thyroid 2003;13:1069–78.
[25] Rønjom MF, Brink C, Lorenzen EL, Hegedüs L, Johansen J. Variation of normal tissue complication probability (NTCP) estimates of radiation-induced hypothyroidism in relation to changes in delineation of the thyroid gland. Acta Oncol Epub 2015 Jan 18.

Supplementary material available online

Supplementary Figure 1 online at: http://informahealthcare.com/doi/abs/10.3109/0284186X.2015.1064160