External validation of an NTCP model for acute esophageal toxicity in locally advanced NSCLC patients treated with intensity-modulated (chemo-)radiotherapy

Radiotherapy and Oncology 129 (2018) 249–256

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Lung cancer

External validation of an NTCP model for acute esophageal toxicity in locally advanced NSCLC patients treated with intensity-modulated (chemo-)radiotherapy

Frank J.W.M. Dankers [a,f,1,*], Robin Wijsman [a,g,1], Esther G.C. Troost [b,c,d,e], Caroline J.A. Tissing-Tan [h], Margriet H. Kwint [i], José Belderbos [i], Dirk de Ruysscher [f], Lizza E. Hendriks [j], Lioe-Fee de Geus-Oei [k,l], Laura Rodwell [m], Andre Dekker [f], René Monshouwer [a], Aswin L. Hoffmann [b,c,d], Johan Bussink [a]

[a] Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands; [b] Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Institute of Radiooncology – OncoRay; [c] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden; [d] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden – Rossendorf; [e] German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany; [f] Department of Radiation Oncology (MAASTRO), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center; [g] Department of Radiation Oncology, University of Groningen, University Medical Center Groningen; [h] Department of Radiation Oncology, Radiotherapiegroep, Arnhem; [i] Department of Radiation Oncology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam; [j] Department of Pulmonary Diseases, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre; [k] Department of Radiology, Leiden University Medical Center; [l] Biomedical Photonic Imaging Group, MIRA Institute, University of Twente, Enschede; and [m] Department for Health Evidence, Radboud University Medical Center, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 8 March 2018
Received in revised form 23 June 2018
Accepted 23 July 2018
Available online 18 September 2018

Keywords:
Non-small cell lung cancer
Acute esophagitis
Intensity-modulated radiation therapy
Predictive modeling
External validation

ABSTRACT

Background and purpose: We externally validated a previously established multivariable normal-tissue complication probability (NTCP) model for Grade ≥2 acute esophageal toxicity (AET) after intensity-modulated (chemo-)radiotherapy or volumetric-modulated arc therapy for locally advanced non-small cell lung cancer.

Materials and methods: A total of 603 patients from five cohorts (A–E) within four different Dutch institutes were included. Using the NTCP model, containing predictors concurrent chemoradiotherapy, mean esophageal dose, gender and clinical tumor stage, the risk of Grade ≥2 AET was estimated per patient and model discrimination and (re)calibration performance were evaluated.

Results: Four validation cohorts (A, B, D, E) experienced higher incidence of Grade ≥2 AET compared to the training cohort (49.3–70.2% *vs* 35.6%; borderline significant for one cohort, highly significant for three cohorts). Cohort C experienced lower Grade ≥2 AET incidence (21.7%, $p < 0.001$). For three cohorts (A–C), discriminative performance was similar to the training cohort (area under the curve (AUC) 0.81–0.89 *vs* 0.84). In the two remaining cohorts (D–E) the model showed poor discriminative power (AUC 0.64 and 0.63). Reasonable calibration performance was observed in two cohorts (A–B), and recalibration further improved performance in all three cohorts with good discrimination (A–C). Recalibration for the two poorly discriminating cohorts (D–E) did not improve performance.

Conclusions: The NTCP model for AET prediction was successfully validated in three out of five patient cohorts (AUC ≥0.80). The model did not perform well in two cohorts, which included patients receiving substantially different treatment. Before applying the model in clinical practice, validation of discrimination and (re)calibration performance in a local cohort is recommended.

 Radiotherapy and Oncology 129 (2018) 249–256

Acute esophageal toxicity (AET) is frequently observed in locally advanced non-small cell lung cancer (LA-NSCLC) patients undergoing (chemo-)radiotherapy, particularly when patients receive concurrent chemotherapy [1,2]. Normal-tissue complication probability (NTCP) models can help to estimate the risk of moderate or severe AET, which may be of benefit for anticipating events of hospitalization or treatment interruptions due to AET [3–7]. These multivariable NTCP models may also be used by doctors as a tool to support their decision on whether or not to treat at the cost of more AET [8–10]. Furthermore, in case there is an increased risk of AET, patients may be selected that benefit most from other radiotherapy techniques such as proton therapy [11,12].

* Corresponding author at: Department of Radiation Oncology 874, Radboud University Medical Center, P.O. Box 9101, Nijmegen 6500 HB, The Netherlands.
E-mail address: frank.dankers@radboudumc.nl (F.J.W.M. Dankers).
[1] Contributed equally.

https://doi.org/10.1016/j.radonc.2018.07.021
0167-8140/

The vast majority of the reported NTCP models for AET are based on 3-dimensional conformal radiotherapy (3D-CRT) techniques. Intensity-modulated radiation therapy (IMRT) and volumetric-modulated arc therapy (VMAT), however, produce more conformal dose distributions at the cost of increased volumes receiving lower dose [13–16]. These differences may result in a different toxicity profile and thus require new NTCP models [17–19]. Therefore, the available NTCP models based on 3D-CRT may not be appropriate for AET risk prediction in patients treated with modern dose delivery techniques. We previously reported on an IMRT- and VMAT-based multivariable NTCP model for Grade ≥2 AET [20]. This model was internally validated and the area under the receiver operating curve (AUC) was 0.84 (0.82 after correction for optimism) indicating good discriminative power of the model. Nonetheless, as reproducibility (model performance on new samples from the same target population), and transportability (model performance on samples from different but related populations) of well internally validated prediction models can still be poor, external validation is needed to assess 'generalizability' of the NTCP model to external patient cohorts [21–24].

In this study, we used five patient cohorts from four different Dutch institutes to externally validate the previously reported multivariable NTCP model for Grade ≥2 AET after IMRT or VMAT for LA-NSCLC (TRIPOD statement Type 4 external validation study [24]).

Materials and methods

Established NTCP model for AET

The model was developed using a training cohort of 149 LA-NSCLC patients who underwent (chemo-)radiotherapy using IMRT or VMAT at the Radboud University Medical Center (Nijmegen, The Netherlands) between March 2008 and June 2013. Information on treatment and patient selection has been previously described in more detail [20]. In brief, all patients received ≥60 Gy (median 66 Gy) in 2 Gy fractions (once daily), with or without (concurrent or sequential) chemotherapy (Table 1). The sequential chemotherapy regimen typically consisted of 3 (3-weekly) courses of gemcitabine/cisplatin, whereas all patients undergoing concurrent chemoradiotherapy (CCR) received 2 (3-weekly) courses of etoposide/cisplatin.

AET was scored weekly during treatment by the treating radiation oncologist using the Radiation Therapy Oncology Group (RTOG) acute radiation morbidity scoring criteria [25]. Toxicity scoring was continued after treatment until acute toxicity resolved. The AET scores were analyzed in relation to clinical risk factors and radiation treatment plan derived dose volume histogram (DVH) parameters.

After multivariable logistic regression, with bootstrap sampling for model order and predictor selection, the following optimal NTCP model for Grade ≥2 AET (maximum at any timepoint) was established:

$$NTCP\left(\underline{x}\right) = \frac{1}{1 + e^{-S\left(\underline{x}\right)}} \tag{1}$$

with,

$$S\left(\underline{x}\right) = -6.418 + 2.645 \cdot CCR + 0.117 \cdot MED + 1.204 \cdot Gender + 0.994 * cT \tag{2}$$

and CCR = concurrent chemoradiotherapy (1 = yes, 0 = no), MED = mean esophageal dose (preferably first converting physical dose to linear-quadratic equivalent dose in 2 Gy fractions with α/β = 10 Gy using MED and its standard deviation [8,26], or esophageal DVH or full dose matrix [27,28]), gender (1 = female, 0 = male) and *c*T = clinical tumor stage (0 < *c*T3, 1 ≥ *c*T3).

External validation cohorts

Five cohorts from four different Dutch institutes were available for validation of the abovementioned NTCP model. The patient, tumor and treatment characteristics of each cohort are listed in Table 1 and Supplementary Material Table S1. Except for cohorts D and E, acute toxicity was retrieved retrospectively for these cohorts from the electronic health records. For all cohorts toxicity was scored weekly during radiotherapy and continued after radiotherapy until toxicity resolved, maximum AET score was used as outcome for model performance evaluation.

Cohort A (n = 47) was also treated in the Department of Radiation Oncology of the Radboud University Medical Center [20]. This cohort consisted solely of stage III NSCLC patients that were treated with (chemo-)radiotherapy using VMAT between June 2013 and December 2014. Radiotherapy and chemotherapy regimens and AET scoring were similar to those of the training cohort. Cohort B (n = 73) consisted of stage III NSCLC patients which received (chemo-)radiotherapy at 'Radiotherapiegroep' (Arnhem, The Netherlands) between January 2014 and March 2016 using mostly VMAT. The radiotherapy regimen and AET scoring were similar to the training cohort. Sequential chemotherapy was platinum based, preferentially cisplatin. Concurrent chemotherapy consisted of 2 courses of platinum/etoposide sometimes preceded by one course of a platinum doublet with either etoposide, or pemetrexed.

Cohort C consisted of 156 stage I-III NSCLC patients treated with (chemo-)radiotherapy at The Netherlands Cancer Institute (Amsterdam, The Netherlands) between December 1998 and March 2003 using 3D-CRT [29]. For 27 patients, however, the predictor 'clinical T-stage' required in the NTCP-model was not available and therefore 129 patients with complete data were included. Varying radiotherapy schedules (total dose 49.5–94.5 Gy, 2.25–2.75 Gy per fraction) were administered, and sequential and concurrent chemotherapy consisted of 2 courses of gemcitabine/cisplatin or daily low-dose cisplatin, respectively. The incidence of AET in this cohort has been evaluated and reported previously; AET was scored using the RTOG scoring criteria [29]. Cohort D was also retrieved from The Netherlands Cancer Institute comprising 172 patients treated between January 2008 and November 2010, and their AET was scored using the Common Toxicity Criteria Adverse Effects (CTCAE) v3.0 [30]. See Table S2 in the Supplementary Material for a comparison between AET scoring using RTOG, CTCAE v3.0 and v4.0. These patients all underwent concurrent chemoradiotherapy (daily low-dose cisplatin) using IMRT (66 Gy in 24 fractions) [31].

The patients from cohort E (n = 398) were treated at MAASTRO Clinic (Maastricht, The Netherlands) between April 2006 and October 2013. Of these, 216 patients had missing data, *i.e.*, missing mean esophageal dose (n = 201, for technical reasons), AET score (n = 4; CTCAE v3.0 and v4.0 [32]), chemotherapy sequence (n = 1) and clinical T-stage (n = 10), and thus 182 patients were included. Patients received 1–3 courses of induction chemotherapy (gemcitabine or cisplatinum) typically followed by concurrent chemotherapy (n = 156) or sequential chemotherapy (n = 24) consisting of 2 courses of a platinum-based doublet. Two patients received no chemotherapy at all. The majority of patients (n = 161) received a total radiation dose of 69 Gy in 1.5 Gy fractions twice daily up to 45 Gy, followed by 8 to 24 Gy in 2 Gy once daily fractions, depending on the dose to the organs at risk (OAR) [33]. Eighteen patients were treated within the FDG-PET-based international multicenter Phase II dose escalation trial "PET-boost" [34]; they received 66 Gy in 24 once daily fractions to the gross tumor

Table 1
Comparison of validation cohort characteristics with the training cohort for the NTCP model predictors and AET.

NTCP model predictors	Training cohort	Validation cohorts											
		A		B		C		D		E		A–E[†]	
	n = 149	*n* = 47	*p*	*n* = 73	*p*	*n* = 129	*p*	*n* = 172	*p*	*n* = 182	*p*	*n* = 603	*p*
Gender (%)													
Male	97 (65.1)	18 (38.3)		38 (52.1)		88 (68.2)		102 (59.3)		113 (62.1)		359 (59.5)	
Female	52 (34.9)	29 (61.7)	**0.002**	35 (47.9)	0.08	41 (31.8)	0.61	70 (40.7)	0.30	69 (37.9)	0.65	244 (40.5)	0.65
T-stage (%)													
≤2	75 (50.3)	21 (44.7)		24 (32.9)		62 (48.1)		91 (52.9)		78 (42.9)		276 (45.8)	
≥3	74 (49.7)	26 (55.3)	0.51	49 (67.1)	**0.02**	67 (51.9)	0.72	81 (47.1)	0.66	104 (57.1)	0.19	327 (54.2)	0.19
Chemotherapy (%)													
Concurrent	93 (62.4)	33 (70.2)		45 (61.6)		25 (19.4)		172 (100.0)		156 (85.7)		431 (71.5)	
Sequential/none	46/10 (37.6)	12/2 (29.8)	0.38	24/4 (38.4)	1.00	31/73 (80.6)	**<0.001**	–	**<0.001**	24/2 (14.3)	**<0.001**	91/81 (28.5)	**<0.001**
D_{mean} esophagus in Gy													
Median physical dose (IQR)	25.2	28.8		26.5		–		–		20.0		25.5	
	(20.5–31.0)	(22.2–34.1)	0.06	(23.3–32.7)	0.16	–		–		(14.9–27.9)	**<0.001**	(17.5–32.7)	0.72
Median $EQD2_{\alpha/\beta=10}$ (IQR)	24.0	–		–		24.1		30.1		–		–	
	(19.6–30.1)	–		–		(10.6–33.3)	0.20	(23.7–36.5)	**<0.001**	–		–	
Grade ≥2 AET													
RTOG	53 (35.6)	33 (70.2)	**<0.001**	36 (49.3)	0.06	28 (21.7)	**0.01**	–		–		323 (53.6)	**<0.001**
CTCAE*	–	–		62 (84.9)	**<0.001**	–		102 (59.3)	**<0.001**	124 (68.1)	**<0.001**	–	
Grade ≥3 AET													
RTOG	13 (8.7)	10 (21.3)	**0.03**	12 (16.4)	0.11	7 (5.4)	0.36	–		–		124 (20.6)	**<0.001**
CTCAE*	–	–		13 (17.8)	0.07	–		40 (23.3)	**<0.001**	55 (30.2)	**<0.001**	–	

Abbreviations: NTCP = normal-tissue complication probability; AET = acute esophageal toxicity; D_{mean} = mean dose; IQR = interquartile range; $EQD2_{10}$ = equivalent dose in 2 Gy fractions with α/β = 10 Gy; RTOG = Radiation Therapy Oncology Group; CTCAE = Common Toxicity Criteria Adverse Effects; N/A = not applicable.
The *p*-values are calculated for the comparison between the validation cohort and the training cohort (Mann–Whitney-U or Fisher's exact test where appropriate). Bold *p*-values are statistically significant. *p-values of AET scoring using CTCAE are calculated with respect to the training cohort AET scoring that used RTOG.
[†] The combined cohort A-E has a mixture of physical and equivalent mean easophageal dose, and a mixture of RTOG and CTCAE-based toxicity scores.

volume (GTV). In case dose escalation was possible (by increasing the fraction dose with equal number of fractions), an integrated boost was delivered to the primary tumor as a whole or to the volume of the primary tumor encompassed by 50% of the maximum standardized uptake value of FDG.

Statistical analysis

Differences between the training cohort from which the NTCP model was developed and the validation cohorts were tested for statistical significance using the Mann–Whitney–U or Fisher's exact test, where appropriate (SPSS software, version 22.0; SPSS Inc., Chicago, USA). A p-value of <0.05 was considered statistically significant.

Model performance

The risk of Grade ≥2 AET was calculated for each individual patient by applying the original NTCP model (Formula 1 and 2). The discriminative power of the model for the validation cohorts was assessed by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC). The criterion for successful external validation was AUC ≥0.80, *i.e.*, no significant deterioration of model performance with respect to the training cohort (AUC 0.84, or 0.82 after optimism correction [20]). Furthermore, the discrimination slopes were calculated by the absolute difference between the mean predicted risk of the groups with and without Grade ≥2 AET.

Model calibration performance was assessed by calibration plots displaying grouped observed frequencies versus predicted outcome [35]. A loess smoother was plotted, which approximates the $y = x$ identity line in case of good calibration [36]. The 95% confidence intervals of the binomially distributed grouped frequencies were calculated according to the Wilson interval [37]. Double histograms of predicted probabilities for patients with and without Grade ≥2 AET were also generated for the calibration plots.

To assess possible miscalibration in the cohorts, the method of logistic recalibration was applied [38,39]. The linear predictors for each patient, *i.e.*, the calculated results after inserting patient specific parameters into Formula 2, were used as a single predictor in a new logistic regression model according to:

$$NTCP\left(\underline{x}\right) = \frac{1}{1 + e^{-S'\left(\underline{x}\right)}} \tag{3}$$

with updated linear predictor

$$S'\left(\underline{x}\right) = a + b \cdot S\left(\underline{x}\right) \tag{4}$$

The resulting calibration intercept a ('calibration-in-the-large') compares the mean of the predicted risks with the mean of the observed risk and gives an indication whether predictions are systematically under- ($a > 0$) or overestimated ($a < 0$). The calibration slope b indicates the level of overfitting ($b < 1$), *i.e.*, the predictions are too extreme, or underfitting ($b > 1$), the predictions are too mild. Recalibration does neither affect sensitivity nor specificity and thus ROC and AUC both remain the same [21,35].

The overall performance of the recalibrated models in each cohort was additionally assessed by calculation of the scaled Brier score, a quadratic scoring rule corrected for dependence on the incidence of the outcome [21]. Additionally, Nagelkerke's R^2 was calculated, which is a logarithmic scoring rule to express the amount of variance in the dependent variables explained by the model [39,40].

Results

Comparison of cohorts

A comparison of training and validation cohort characteristics for the NTCP model predictors and AET is listed in Table 1. The incidence of Grade ≥2 AET in cohorts A, D and E was (nearly) twice the incidence of Grade ≥2 AET in the training cohort (70.2%, 59.3% and 68.1% *vs* 35.6%, respectively; $p < 0.001$). The patients in cohort C experienced lower rates of Grade ≥2 AET compared to the training cohort (21.7% *vs* 35.6%, respectively; $p = 0.01$). Other patient, tumor and treatment characteristics of the cohorts are listed as Supplementary Material in Table S1.

Model performance

A summary of model performance in the validation cohorts, *i.e.*, overall performance, discrimination and (re)calibration, is listed in Table 2. Unsurprisingly, the best performance, as indicated by the highest value of the scaled Brier and Nagelkerke R^2, was seen in the training cohort. The overall performance was high for cohorts A, B and C, but was poor for cohorts D and E.

The ROC curves for all cohorts are displayed in Fig. 1. High discriminative performance of similar quality to the training cohort was obtained for cohorts A, B and C, as indicated by high AUCs (0.89, 0.81 and 0.84, respectively). Poor discrimination of the model was found in cohorts D and E (AUC 0.64 and 0.63 respectively). This poor discrimination performance is also demonstrated by the calculated discrimination slopes (Table 2).

Model calibration performance, without recalibration, can be visually assessed from the calibration plots shown in Fig. S1 of the Supplementary Material. Reasonable performance without

Table 2
Performance of the NTCP model after recalibration for the different patient cohorts.

Performance measure	Training cohort	Validation cohort					
		A	B	C	D	E	A–E
	$n = 149$	$n = 47$	$n = 73$	$n = 129$	$n = 172$	$n = 182$	$n = 603$
Pseudo R^2s							
$\text{Brier}_{\text{scaled}}$	0.35	0.44	0.31	0.24	0.06	0.05	0.19
Nagelkerke	0.41	0.55	0.38	0.36	0.08	0.06	0.24
Discrimination							
AUC (95% CI)	0.84 (0.77–0.91)	0.89 (0.80–0.98)	0.81 (0.70–0.91)	0.84 (0.75–0.94)	0.64 (0.55–0.72)	0.63 (0.55–0.71)	0.74 (0.70–0.78)
SE	0.04	0.05	0.05	0.05	0.04	0.04	0.02
Discrimination slope	0.33	0.45	0.30	0.25	0.06	0.05	0.19
Calibration							
Calibration-in-the-large	0.00	1.18	0.20	−0.15	−0.22	1.63	0.57
Calibration slope	1.00	1.36	0.71	0.60	0.40	0.29	0.50

Abbreviations: NTCP = normal-tissue complication probability; AUC = area under the curve; CI = confidence interval; SE = standard error.

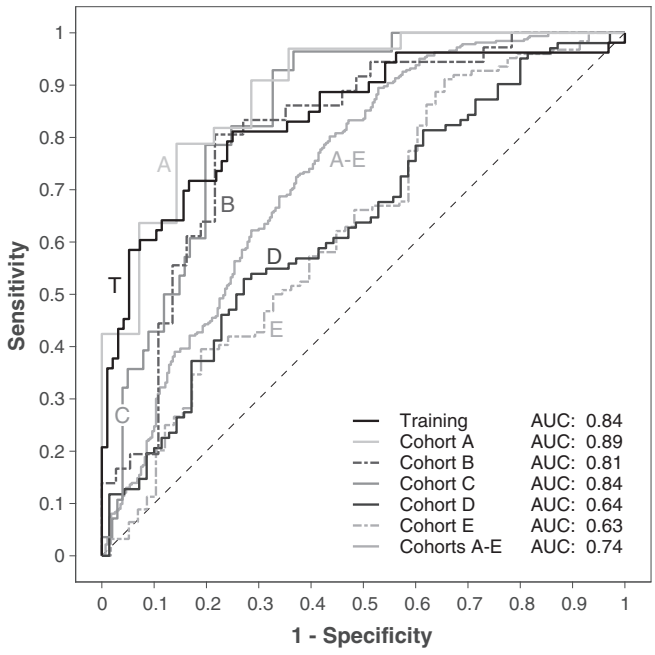

Fig. 1. ROC curves of the previously published NTCP model [20] applied on all patient cohorts showing good discriminating performance for 3 out of 5 validation cohorts as indicated by AUC values (>0.80). *Abbreviations*: ROC = receiver operating characteristic; NTCP = normal-tissue complication probability; AUC = area under the curve.

recalibration was found for the model in cohorts A and B, demonstrated by the loess smoother which was relatively close to the identity line. The model generally underestimated the risk of Grade ≥2 AET. Increasingly poor calibration was observed for cohorts C, D and E.

Calibration plots generated after recalibration are shown in Fig. 2, and the values for the calibration-in-the-large and calibration slope are listed in Table 2. For cohorts A and B, good calibration was achieved after recalibration. Similarly, for cohort C recalibration moderately improved the agreement between predicted and observed risk. For cohorts D and E, calibration did not improve after recalibration, indicated by the limited range of predicted probabilities (see Fig. 2).

Discussion

Recently, we established a multivariable NTCP model for AET in LA-NSCLC undergoing IMRT or VMAT and after thorough internal validation the model proved to be robust [20]. However, it is of paramount importance to perform external validation in order to ensure that the model is transportable to other patient cohorts [21,23]. This means that the model produces accurate predictions in a sample that was drawn from a different but plausibly related population. Several components of 'transportability' can be distinguished, such as historical (*e.g.*, a different time period), geographical (*e.g.*, treated in a different hospital) and methodological (*e.g.*, differences in toxicity scoring) transportability [41]. To account for all these components of transportability, we externally validated our previously established NTCP model for Grade ≥2 AET in cohorts of (LA-)NSCLC patients that were treated by (chemo-) radiotherapy in different hospitals (cohorts B-E), receiving different radiation fractionation schedules (cohorts C-E) and in a historically different period of time with less conformal dose delivery techniques (cohort C). Ideally, an NTCP model performs well in every patient cohort external to the cohort the model was developed on. However, this so-called 'strong calibration' is only considered possible in utopia [35]. Therefore, applying an established NTCP model in different patient cohorts often needs some form of adjustments to account for local circumstances [42,43].

Recalibration is a controlled form of model updating; *i.e.*, the coefficients of the model are adjusted to correct for differences in for instance event rates. Initial calibration of the model in cohorts A and B was moderate (see Fig. S1 in the Supplementary Material). Underestimation of Grade ≥2 AET was seen, which is possibly due to a lower incidence of Grade ≥2 AET in the training cohort (35.6%) compared to cohort A (70.2%) and cohort B (49.3%). The class imbalance in the training cohort can affect the estimate of the model intercept and skews the predicted probabilities. After recalibration of the NTCP model for cohorts A and B, calibration improved (see Fig. 2). Discrimination of the model was good for the patients in cohorts A and B (AUC 0.89 and 0.81, respectively). Formerly, we hypothesized that differences in dose delivery techniques influenced NTCP modeling since the models based on 3D-CRT did not perform well in head and neck cancer patients who underwent IMRT [18,20,44,45]. Although cohort C differs substantially from the training cohort regarding treatment technique (3D-CRT *vs* IMRT/VMAT), radiation dose (49.5–94.5 Gy *vs* 66 Gy), the application of concurrent chemotherapy, and the time period (1998–2003 *vs* 2008–2010), the current model performed surprisingly well for this population (AUC 0.84 with a moderately good recalibration curve). Cohorts D and E showed poor discrimination (AUC 0.64 and 0.63 respectively) and (re)calibration (see Fig. 1 and Supplementary Material Fig. S1). Re-estimating the regression coefficients or adding additional predictors that are known for their association with AET (for example, overall treatment time (OTT) and chemotherapy regimen; see below) are approaches to improve model predictions. Besides this, there may be several other reasons for the poor model performance in these cohorts. Firstly, the NTCP model was developed using the RTOG grading scale for AET. However, toxicity for the patients in cohorts D and E was scored using the CTCAE grading scales for AET. Differences between scoring systems were reported to be of importance in modeling of toxicity, for instance for modeling the risk of radiation-induced pneumonitis [46]. It is likely that such differences in grading scales affect AET modeling as well. This was illustrated for the patients of cohort B for whom both the RTOG and CTCAE v4.0 grading of AET were available. Applying the NTCP model using the CTCAE-based AET scores resulted in a high discrimination with AUC of 0.80 (compared to 0.81 for the RTOG based scores), however, model calibration was poor since it considerably underestimated the risk of CTCAE Grade ≥2 AET (data not shown). The latter can be explained by the finding that in 35.6% of the patients AET was scored as Grade 1 using the RTOG scale and as Grade 2 using the CTCAE scale (see Table S2 in the Supplementary Material). Secondly, the patients from cohort D received concurrent chemoradiotherapy in a fundamentally different protocol compared to the patients in the training cohort as they received daily low-dose cisplatin and moderately hypofractionated radiotherapy schedules. Thirdly, the OTT is shorter for cohorts D and E (5 weeks) than for the training cohort (6.5 weeks). Besides, the majority of patients (88.5%) from cohort E were treated twice-daily. Both factors are known to result in a strong increase of AET [3,6]; including OTT in the NTCP model for patients receiving treatment with a shorter OTT is likely to improve model performance for these cohorts as reported by Dehing-Oberije et al. [3].

Despite our aim to thoroughly validate the established NTCP model for Grade ≥2 AET by assessing the transportability of the model using multiple different patient cohorts, some potential limitations should be noted. Firstly, the data of most cohorts were retrieved retrospectively (except cohorts D and E) possibly introducing unwanted bias. Furthermore, for some patients of the validation cohorts the necessary NTCP model predictor values could

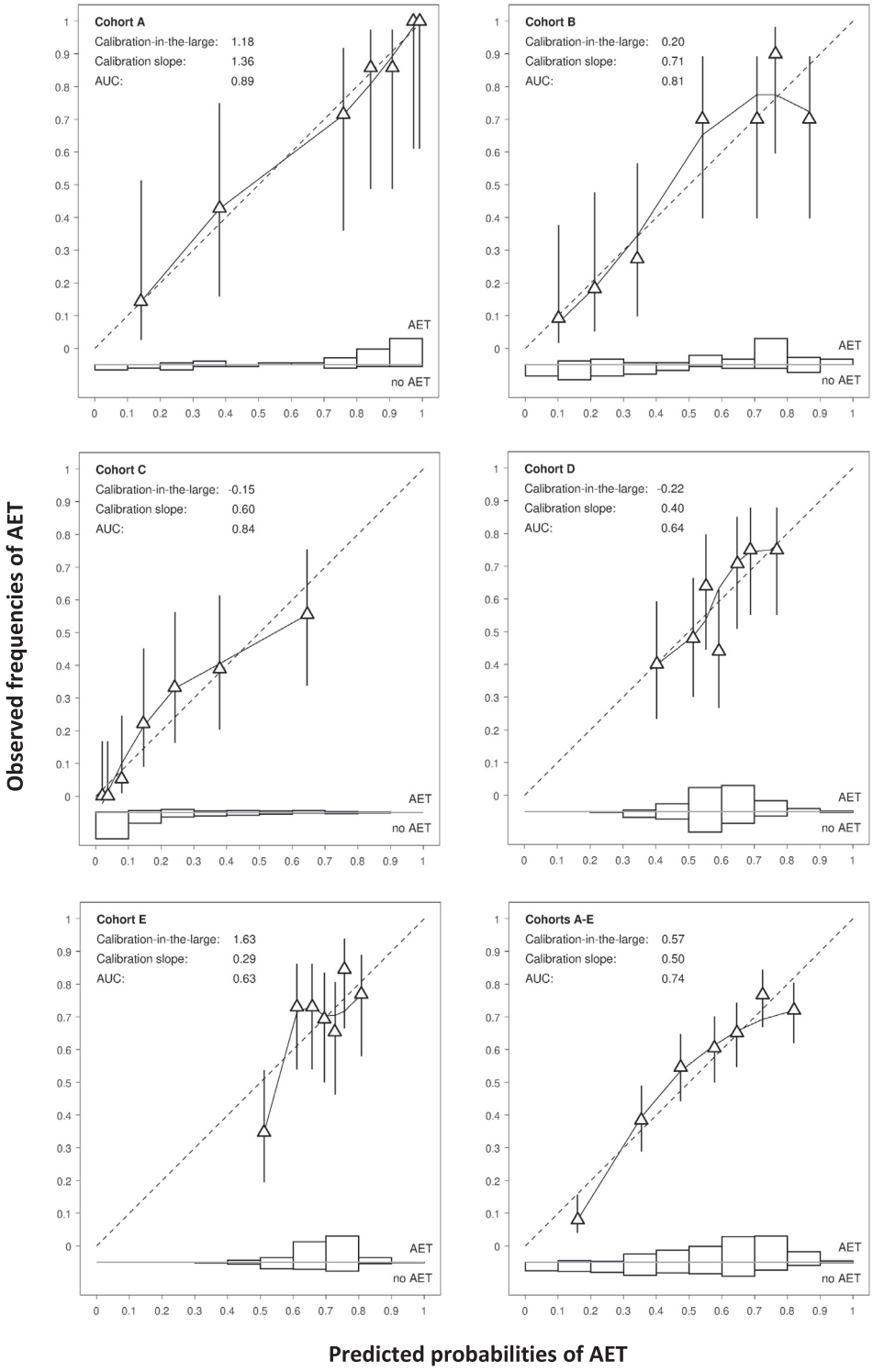

Fig. 2. Calibration plots of the NTCP model applied on all validation cohorts separate and combined, after recalibration per cohort. Recalibrated predicted probabilities are calculated by inserting the cohort-specific calibration-in-the-large and calibration slope values in Formulas 3 and 4. The triangles indicate grouped predicted probabilities of Grade ≥2 AET *vs* grouped observed frequencies. The vertical lines represent 95% confidence intervals. A loess smoother was fitted and displayed by the black line. Perfect predictions should be close to the dashed 45° reference line. Double histograms of patients with and without Grade ≥2 AET, binned according to their predicted probabilities, are displayed at the bottom. *Abbreviations*: NTCP = normal-tissue complication probability; AET = acute esophageal toxicity; AUC = area under the curve.

not be retrieved resulting in exclusion of these patients. The number of patients of the separate cohorts may be considered low for model validation, however, the total number of patients (n = 603) included in the validation cohorts is substantial. For future work, by making data 'smarter', *e.g.*, by implementing semantic technologies [47,48], and more easily accessible, by adhering to the FAIR data principles [49], distributed learning techniques can allow training and validation of models in much larger cohorts of patients that were not treated according to any specific study protocol [50]. Finally, this study is an external validation of a model previously published by us and we therefore encourage independent external validation by other research groups.

In conclusion, the established NTCP model for the prediction of Grade ≥2 AET in patients treated for locally advanced NSCLC successfully validated in 3 out of 5 patient cohorts, but performed poor in 2 cohorts that were significantly different for many variables. Before implementing the NTCP model in clinical practice, one should always check model discrimination and calibration performance in a local cohort representative of the patients for which the model is intended to be used in the future. If good discrimination but poor calibration is observed a local recalibration of the model is advised. After implementation the model should be evaluated over time for new patients since treatments and cohorts change and model performance can deteriorate to the point where the model coefficients need to be updated or additional predictors may become relevant and complete remodeling is necessary.

Conflict of interests

Andre Dekker is a founder and shareholder of Medical Data Works B.V. which provides services for prediction modeling. MAASTRO Clinic receives research funding from Varian Medical Systems for prediction modeling research. Lizza Hendriks reports personal fees from Roche, MSD, AstraZeneca and BMS, all outside the scope of this submitted research.

Acknowledgments

The authors thank Kasper Pasma (Radiotherapiegroep, Arnhem, The Netherlands) for assistance in providing validation cohort data and Ton de Haan (Radboud University Medical Center, Nijmegen, The Netherlands) for valuable input in conducting the statistical analyses.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.radonc.2018.07.021.

References

[1] Auperin A, Le Pechoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol 2010;28:2181–90.
[2] Palma DA, Senan S, Oberije C, Belderbos J, de Dios NR, Bradley JD, et al. Predicting esophagitis after chemoradiation therapy for non-small cell lung cancer: an individual patient data meta-analysis. Int J Radiat Oncol Biol Phys 2013;87:690–6.
[3] Dehing-Oberije C, De Ruysscher D, Petit S, Van Meerbeeck J, Vandecasteele K, De Neve W, et al. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol 2010;97:455–61.
[4] Huang EX, Bradley JD, El Naqa I, Hope AJ, Lindsay PE, Bosch WR, et al. Modeling the risk of radiation-induced acute esophagitis for combined Washington University and RTOG trial 93–11 lung cancer patients. Int J Radiat Oncol Biol Phys 2012;82:1674–9.
[5] Zhang ZC, Xu J, Li BS, Zhou T, Lu J, Wang ZT, et al. Clinical and dosimetric risk factors of acute esophagitis in patients treated with 3-dimensional conformal radiotherapy for non-small-cell lung cancer. Am J Clin Oncol 2010;33:271–5.
[6] Werner-Wasik M, Yorke E, Deasy J, Nam J, Marks LB. Radiation dose-volume effects in the esophagus. Int J Radiat Oncol Biol Phys 2010;76:S86–93.
[7] Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. Radiother Oncol 2014;112:37–43.
[8] Hoffmann AL, Troost EG, Huizenga H, Kaanders JH, Bussink J. Individualized dose prescription for hypofractionation in advanced non-small-cell lung cancer radiotherapy: an in silico trial. Int J Radiat Oncol Biol Phys 2012;83:1596–602.
[9] Bradley JD, Paulus R, Komaki R, Masters G, Blumenschein G, Schild S, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. Lancet Oncol 2015;16:187–99.
[10] Even AJ, van der Stoep J, Zegers CM, Reymen B, Troost EG, Lambin P, et al. PET-based dose painting in non-small cell lung cancer: Comparing uniform dose escalation with boosting hypoxic and metabolically active sub-volumes. Radiother Oncol 2015;116:281–6.
[11] 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 2013;107:267–73.
[12] Widder J, van der Schaaf A, Lambin P, Marijnen CA, Pignol JP, Rasch CR, et al. The quest for evidence for proton therapy: model-based approach and precision medicine. Int J Radiat Oncol Biol Phys 2016;95:30–6.
[13] Jiang X, Li T, Liu Y, Zhou L, Xu Y, Zhou X, et al. Planning analysis for locally advanced lung cancer: dosimetric and efficiency comparisons between intensity-modulated radiotherapy (IMRT), single-arc/partial-arc volumetric modulated arc therapy (SA/PA-VMAT). Radiat Oncol 2011;6:140–7.
[14] Christian JA, Bedford JL, Webb S, Brada M. Comparison of inverse-planned three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for non-small-cell lung cancer. Int J Radiat Oncol Biol Phys 2007;67:735–41.
[15] Wijsman R, Dankers F, Troost EG, Hoffmann AL, van der Heijden EH, de Geus-Oei LF, et al. Comparison of toxicity and outcome in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-) radiotherapy using IMRT or VMAT. Radiother Oncol 2017;122:295–9.
[16] Monshouwer R, Hoffmann AL, Kunze-Busch M, Bussink J, Kaanders JH, Huizenga H. A practical approach to assess clinical planning tradeoffs in the design of individualized IMRT treatment plans. Radiother Oncol 2010;97:561–6.
[17] Gomez DR, Tucker SL, Martel MK, Mohan R, Balter PA, Lopez Guerra JL, et al. Predictors of high-grade esophagitis after definitive three-dimensional conformal therapy, intensity-modulated radiation therapy, or proton beam therapy for non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2012;84:1010–6.
[18] Beetz I, Schilstra C, van Luijk P, Christianen ME, Doornaert P, Bijl HP, et al. External validation of three dimensional conformal radiotherapy based NTCP models for patient-rated xerostomia and sticky saliva among patients treated with intensity modulated radiotherapy. Radiother Oncol 2012;105:94–100.
[19] Dankers F, Wijsman R, Troost EG, Monshouwer R, Bussink J, Hoffmann AL. Esophageal wall dose-surface maps do not improve the predictive performance of a multivariable NTCP model for acute esophageal toxicity in advanced stage NSCLC patients treated with intensity-modulated (chemo-) radiotherapy. Phys Med Biol 2017;62:3668–81.
[20] Wijsman R, Dankers F, Troost EG, Hoffmann AL, van der Heijden EH, de Geus-Oei LF, et al. Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-)radiotherapy. Radiother Oncol 2015;117:49–54.
[21] 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;21:128–38.
[22] Yahya N, Ebert MA, Bulsara M, Kennedy A, Joseph DJ, Denham JW. Independent external validation of predictive models for urinary dysfunction following external beam radiotherapy of the prostate: Issues in model development and reporting. Radiother Oncol 2016;120:339–45.
[23] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279–89.
[24] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
[25] Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys 1995;31:1341–6.
[26] Hoffmann AL, Nahum AE. Fractionation in normal tissues: the (alpha/beta)eff concept can account for dose heterogeneity and volume effects. Phys Med Biol 2013;58:6897–914.
[27] Wheldon TE, Deehan C, Wheldon EG, Barrett A. The linear-quadratic transformation of dose-volume histograms in fractionated radiotherapy. Radiother Oncol 1998;46:285–95.
[28] Bentzen SM, Dorr W, Gahbauer R, Howell RW, Joiner MC, Jones B, et al. Bioeffect modeling and equieffective dose concepts in radiation oncology–terminology, quantities and units. Radiother Oncol 2012;105:266–8.
[29] Belderbos J, Heemsbergen W, Hoogeman M, Pengel K, Rossi M, Lebesque J. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. Radiother Oncol 2005;75:157–64.
[30] CTCAE v3.0: Common Terminology Criteria for Adverse Events v3.0. National Cancer Institute, DCTD, NCI, NIH, DHHS; March 31, 2003.
[31] Kwint M, Uyterlinde W, Nijkamp J, Chen C, de Bois J, Sonke JJ, et al. Acute esophagus toxicity in lung cancer patients after intensity modulated radiation therapy and concurrent chemotherapy. Int J Radiat Oncol Biol Phys 2012;84:e223–8.
[32] CTCAE v4.0: Common Terminology Criteria for Adverse Events v4.0. National Cancer Institute, NCI, NIH, DHHS (NIH publication # 09-7473); May 29, 2009.
[33] van Baardwijk A, Wanders S, Boersma L, Borger J, Ollers M, Dingemans AM, et al. Mature results of an individualized radiation dose prescription study

based on normal tissue constraints in stages I to III non-small-cell lung cancer. J Clin Oncol 2010;28:1380–6.
[34] van Elmpt W, De Ruysscher D, van der Salm A, Lakeman A, van der Stoep J, Emans D, et al. The PET-boost randomised phase II dose-escalation trial in non-small cell lung cancer. Radiother Oncol 2012;104:67–71.
[35] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76.
[36] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014;33:517–35.
[37] Brown LD, Cai TT, DasGupta A, Agresti A, Coull BA, Casella G, et al. Interval estimation for a binomial proportion – Comment – Rejoinder. Stat Sci 2001;16:101–33.
[38] Cox DR. 2 further applications of a model for binary regression. Biometrika 1958;45:562–5.
[39] Steyerberg EW. Clinical prediction models; Chapter 15. Springer Science +Business Media; 2009.
[40] Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika 1991;78:691–2.
[41] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
[42] Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23:2567–86.
[43] Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth 2009;56:194–201.
[44] Veldeman L, Madani I, Hulstaert F, De Meerleer G, Mareel M, De Neve W. Evidence behind use of intensity-modulated radiotherapy: a systematic review of comparative clinical studies. Lancet Oncol 2008;9:367–75.
[45] Staffurth J, Radiotherapy, Development, Board. A review of the clinical evidence for intensity-modulated radiotherapy. Clin Oncol (R Coll Radiol) 2010;22:643–57.
[46] Tucker SL, Jin H, Wei X, Wang S, Martel MK, Komaki R, et al. Impact of toxicity grade and scoring system on the relationship between mean lung dose and risk of radiation pneumonitis in a large cohort of patients with non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2010;77:691–8.
[47] Traverso A. The Radiation Oncology Ontology (ROO): publishing linked data in radiation oncology using Semantic Web and Ontology techniques. Med Phys 2018. https://doi.org/10.1002/mp.12879 [in press].
[48] Prud'hommeaux E, Seaborne A. SPARQL Query Language for RDF. W3C Recomm; 2008: 15.
[49] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
[50] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9.