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

Citation for published version (APA):

Document status and date:
Published: 01/11/2018

DOI:
10.1016/j.radonc.2018.07.021

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:
- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:
repository@maastrichtuniversity.nl
providing details and we will investigate your claim.

Download date: 28 Mar. 2021
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,⇑, Robin Wijsman a,g,⇑, Esther G.C. Troost b,c,d,e, Caroline J.A. Tissing-Tan h, Margriet H. Kwint i, José Belderbos i, Dirk de Ruyssscher f, Lizza E. Hendriks i, 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 Department of Radiation Oncology and Radiotherapy, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden; c OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden; d German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany; e Department of Radiation Oncology (MAASTRO), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center; f Department of Radiation Oncology, University of Groningen, University Medical Center Groningen; g Department of Radiation Oncology, Radiotherapiegroep, Arnhem; h Department of Radiation Oncology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam; i Department of Pulmonary Diseases, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre; j Biomedical Photonic Imaging Group, MIRA Institute, University of Twente, Enschede; and k,l Department for Health Evidence, Radboud University Medical Center, Nijmegen, The Netherlands

A R T I C L E   I N F O

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

A B S T R A C T

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 discriminative and (re)calibration performance in a local cohort is recommended.

© 2018 Elsevier B.V. All rights reserved. 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].
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:

\[
NTCP(x) = \frac{1}{1 + e^{-S(x)}}
\]

with

\[
S(x) = -6.418 + 2.645 \cdot CCR + 0.117 \cdot MED + 1.204 \cdot Gender + 0.994 \cdot cT
\]

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 \(\alpha/\beta = 10\) Gy using MED and its standard deviation [8,26]), or esophageal DVH or full dose matrix [27,28]), gender (1 = female, 0 = male) and \(cT = \) clinical tumor stage \((0 < cT3, 1 \geq cT3)\).

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 cisplatin) 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>
<thead>
<tr>
<th>NTCP model predictors</th>
<th>Training cohort</th>
<th>Validation cohort A</th>
<th>Validation cohort B</th>
<th>Validation cohort C</th>
<th>Validation cohort D</th>
<th>Validation cohort E</th>
<th>Combined cohort A–E</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 149</td>
<td>n = 47</td>
<td>n = 73</td>
<td>n = 129</td>
<td>n = 172</td>
<td>n = 182</td>
<td>n = 603</td>
<td></td>
</tr>
<tr>
<td>Gender (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>97 (65.1)</td>
<td>18 (38.3)</td>
<td>38 (52.1)</td>
<td>88 (68.2)</td>
<td>102 (59.3)</td>
<td>113 (62.1)</td>
<td>359 (59.5)</td>
</tr>
<tr>
<td>Female</td>
<td>52 (34.9)</td>
<td>29 (61.7)</td>
<td>35 (47.9)</td>
<td>41 (31.8)</td>
<td>70 (40.7)</td>
<td>69 (37.9)</td>
<td>244 (40.5)</td>
</tr>
<tr>
<td>T-stage (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤2</td>
<td>75 (50.3)</td>
<td>21 (44.7)</td>
<td>24 (32.9)</td>
<td>62 (48.1)</td>
<td>91 (52.9)</td>
<td>78 (42.9)</td>
<td>276 (45.8)</td>
</tr>
<tr>
<td>≥3</td>
<td>74 (49.7)</td>
<td>26 (55.3)</td>
<td>49 (67.1)</td>
<td>67 (51.9)</td>
<td>81 (47.1)</td>
<td>104 (57.1)</td>
<td>327 (54.2)</td>
</tr>
<tr>
<td>Chemotherapy (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concurrent</td>
<td>93 (62.4)</td>
<td>33 (70.2)</td>
<td>45 (61.6)</td>
<td>25 (19.4)</td>
<td>172 (100.0)</td>
<td>156 (85.7)</td>
<td>431 (71.5)</td>
</tr>
<tr>
<td>Sequential/none</td>
<td>46/10 (37.6)</td>
<td>12/2 (29.8)</td>
<td>24/4 (38.4)</td>
<td>31/73 (80.6)</td>
<td>&lt;0.001</td>
<td>24/2 (14.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(D_{\text{mean}}) esophagus in Gy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median physical dose (IQR)</td>
<td>25.2 (20.5–31.0)</td>
<td>28.8 (22.2–34.1)</td>
<td>26.5 (23.3–32.7)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Median EQD(_{2,10}) (IQR)</td>
<td>24.0 (19.6–30.1)</td>
<td>–</td>
<td>24.1 (10.6–33.3)</td>
<td>30.1 (23.7–36.5)</td>
<td>&lt;0.001</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Grade ≥2 AET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTOG</td>
<td>53 (35.6)</td>
<td>33 (70.2)</td>
<td>&lt;0.001</td>
<td>36 (49.3)</td>
<td>0.06</td>
<td>28 (21.7)</td>
<td>0.01</td>
</tr>
<tr>
<td>CTCAE(^*)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>62 (84.9)</td>
<td>&lt;0.001</td>
<td>102 (59.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Grade ≥3 AET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTOG</td>
<td>13 (8.7)</td>
<td>10 (21.3)</td>
<td>0.03</td>
<td>12 (16.4)</td>
<td>0.11</td>
<td>7 (5.4)</td>
<td>0.36</td>
</tr>
<tr>
<td>CTCAE(^*)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>13 (17.8)</td>
<td>0.07</td>
<td>40 (23.3)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Abbreviations:** NTCP = normal-tissue complication probability; AET = acute esophageal toxicity; \(D_{\text{mean}}\) = mean dose; IQR = interquartile range; EQD\(_{2,10}\) = equivalent dose in 2 Gy fractions with \(\alpha/\beta = 10\) 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. ^1^ The combined cohort A–E has a mixture of physical and equivalent mean esophageal 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 the training cohort were calculated according to the Wilson interval [37]. Double his-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.

The discriminative performance of the model for 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 shown 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>
<thead>
<tr>
<th>Performance measure</th>
<th>Training cohort</th>
<th>Validation cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>n=149</td>
<td>n=47</td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brief recalibrated</td>
<td>0.35</td>
<td>0.44</td>
</tr>
<tr>
<td>Nagelkerke</td>
<td>0.41</td>
<td>0.55</td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC (95% CI)</td>
<td>0.84 (0.77–0.91)</td>
<td>0.89 (0.80–0.98)</td>
</tr>
<tr>
<td>SE</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Discrimination slope</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>Calibration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration-in-the-large</td>
<td>0.00</td>
<td>1.18</td>
</tr>
<tr>
<td>Calibration slope</td>
<td>1.00</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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.
In conclusion, the established NTCP model for the prediction of Grade ≥2 AET in patients treated for locally advanced NSCLC success-fu-## successfully validated in 3 out of 5 patient cohorts, but performed poor in 2 cohorts that were significantly different for many vari-## ables. 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 discrimin-## ation but poor calibration is observed a local recalibration of the## model is advised. After implementation the model should be eval-## uated 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. Liza 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

[19] CTCAE v4.0: Common Terminology Criteria for Adverse Events v4.0.## National Cancer Institute, NCI, NIH, DHHS (NIH publication # 09-7473); May## 29, 2009.