# The relation between prediction model performance measures and patient selection outcomes for proton therapy in head and neck cancer

Leeuwenberg, Artuur Marijn; Reitsma, Johannes Bernardus; Van den Bosch, Lisa Griet Lydia Jozef; Hoogland, Jeroen; van der Schaaf, Arjen; Hoebers, Frank Jozef Pieter; Wijers, Oda Bemadette; Langendijk, Johannes Albertus; Moons, Karel Gerardus Maria; Schuit, Ewoud

*Published in:*
Radiotherapy and Oncology

*DOI:*
10.1016/j.radonc.2022.109449



*Publication date:*
2023

Radiotherapy and Oncology 179 (2023) 109449

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

# The relation between prediction model performance measures and patient selection outcomes for proton therapy in head and neck cancer

Artuur Marijn Leeuwenberg [a,*], Johannes Bernardus Reitsma [a], Lisa Griet Lydia Jozef Van den Bosch [b], Jeroen Hoogland [a], Arjen van der Schaaf [b], Frank Jozef Pieter Hoebers [c], Oda Bemadette Wijers [d], Johannes Albertus Langendijk [b], Karel Gerardus Maria Moons [a], Ewoud Schuit [a]

[a] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht; [b] Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen; [c] Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht; and [d] Radiotherapeutic Institute Friesland, Leeuwarden, the Netherlands

ARTICLE INFO

*Article history:*
Received 26 October 2022
Received in revised form 8 December 2022
Accepted 13 December 2022
Available online 22 December 2022

*Keywords:*
Prediction performance measures
Normal tissue complication probability models
Head and neck cancer
Individualized treatment decisions

ABSTRACT

*Background:* Normal-tissue complication probability (NTCP) models predict complication risk in patients receiving radiotherapy, considering radiation dose to healthy tissues, and are used to select patients for proton therapy, based on their expected reduction in risk after proton therapy versus photon radiotherapy (ΔNTCP). Recommended model evaluation measures include area under the receiver operating characteristic curve (AUC), overall calibration (CITL), and calibration slope (CS), whose precise relation to patient selection is still unclear. We investigated how each measure relates to patient selection outcomes.
*Methods:* The model validation and consequent patient selection process was simulated within empirical head and neck cancer patient data. By manipulating performance measures independently via model perturbations, the relation between model performance and patient selection was studied.
*Results:* Small reductions in AUC (-0.02) yielded mean changes in ΔNTCP between 0.9–3.2 %, and single-model patient selection differences between 2–19 %. Deviations (-0.2 or +0.2) in CITL or CS yielded mean changes in ΔNTCP between 0.3–1.4 %, and single-model patient selection differences between 1–10 %.
*Conclusions:* Each measure independently impacts ΔNTCP and patient selection and should thus be assessed in a representative sufficiently large external sample. Our suggested practical model selection approach is considering the model with the highest AUC, and recalibrating it if needed.


Normal-tissue complication probability (NTCP) models are prediction models that estimate the risk of experiencing complications after receiving radiotherapy. The risk of complications estimated by an NTCP model is based on patient characteristics, and importantly: predictors taking into account the radiation dosage to healthy (normal) tissue surrounding the tumor. NTCP models have been developed for many complications and tumor regions (e.g., swallowing problems in head and neck cancer patients).[1–6] Besides providing information to patients regarding complication risk, and optimization of radiation treatment plans [7–12], NTCP models have been used to select patients for proton therapy (PT) versus more common photon-based radiotherapy (RT), based on the reduction in predicted complication risk (ΔNTCP) when using PT versus RT.[13–15] The ΔNTCP is calculated by applying the same NTCP model twice for each patient: once with the dose predictor values for the PT treatment plan, and once using the dose predictor values of the RT treatment plan, resulting in a predicted risk per treatment plan. The difference between these risks ($NTCP_{RT}$ - $NTCP_{PT}$) is the ΔNTCP, i.e. the expected reduction in complications because of a reduction in dose to healthy tissues with PT as compared to RT. If the ΔNTCP reaches a certain threshold for a certain complication then PT is considered.

The model-based patient selection procedure (MBS) is used in seven centers in Europe (Denmark, France, Germany, Italy and three in the Netherlands) to select patients for proton therapy for

*Abbreviations:* NTCP, Normal-tissue complication probability; AUC, area under the receiver operating characteristic curve; CITL, calibration in-the-large; CS, Calibration slope; MBS, model-based patient selection; HNC, head- and neck cancer; NIPP, national indication protocol proton therapy.

* Corresponding author at: Str. 6.131, Universiteitsweg 100, 3508 GA Utrecht, the Netherlands.
*E-mail address:* a.m.leeuwenberg-15@umcutrecht.nl (A.M. Leeuwenberg).

https://doi.org/10.1016/j.radonc.2022.109449
0167-8140/

particular tumor sites (e.g., head- and neck cancer, lymphoma, breast cancer, lung cancer).[16] It was firstimplemented in clinical practice in the Netherlands, to select head- and neck cancer (HNC) patients for PT.[17–19] For this, a national indication protocol proton therapy (NIPP) was developed.[18] Currently, NTCP models used in MBS are evaluated via general best practice recommendations for medical prediction model validation: by estimating calibration and discrimination measures.[20–23] In the NIPP, both risk probabilities (NTCP) and risk differences (ΔNTCP) are of interest for clinical decision making, and while the recommended performance measures focus on the first, their influence on the second is less straightforward. Although it is evident that better values for each performance measure indicate better NTCP models, it remains unclear to what degree small changes in prediction performance (calibration and discrimination) influence the actual patient selection. This, for example, complicates the decision to choose between NTCP models that predict the same outcome but perform better at different performance measures. To illustrate this point we provide an example in Explanation Box 1.

The aim of this study is to investigate how these prediction performance measures relate to changes in patient selection in an MBS setting. Knowing this may better inform NTCP model development and validation, geared towards MBS. We investigate this relation via a simulation study of the entire NTCP model validation and MBS deployment process in clinical data.

Box 1 Illustrative imaginary example to demonstrate the challenge of model selection. Imagine that to develop an indication protocol to select patients based on their risk of long-term xerostomia a systematic literature review is performed to identify existing models that predict risk of moderate-to-severe xerostomia six months after receiving radiotherapy. Let us assume three candidate models were identified in the literature, and to compare these models these were externally validated in a new sufficiently large set of patients from the hospital in which the patient selection is intended to be introduced. External validation results of the identified models are shown in Table 1.

It is clear that model B is the least promising model, as it has the worst value for all measures: the lowest area under the receiver operating characteristic curve (AUC), the calibration in-the-large (CITL) deviates most from the ideal value 0, and the calibration slope (CS) is furthest from the ideal value of 1. However, choosing between model A and C is non-trivial, as model A has a better CITL than model C (0.0 versus 0.2) and better CS (1.0 versus 1.2), but model C has a better AUC than model A (0.74 versus 0.71). Without knowing the relation of each measure with consequent patient selection it remains difficult to choose between models A and C.

## Materials and methods

We simulated the model validation and deployment in MBS to inspect how changes in common prediction performance measures (see Explanation Box 1) relate to changes in consequent patient selection. To assess the impact of each *individual* model performance measure on the patient selection outcomes, we varied the NTCP models targeting only one performance measure at the time. Then, using the manipulated models, we performed MBS and observed the changes in patient selection. An overview of the study flow is shown in Fig. 1. The next sections provide details about the used models, manipulations, model selection procedure and the eventual patient selection outcomes.

### *Models*

Four models from the NIPP for model-based selection of HNC patients for proton therapy were used, predicting two severity levels (moderate-to-severe and severe complications) of patient-reported xerostomia, and two severity levels of physician-rated (grade II-V or grade III-V) dysphagia at 6 months after finishing radiotherapy. These NTCP models are multivariable logistic regression models. Their coefficients are reported in Table 2.

### *Study data*

Two empirical datasets were used, one model validation set on which prediction performance measures were calculated to assess

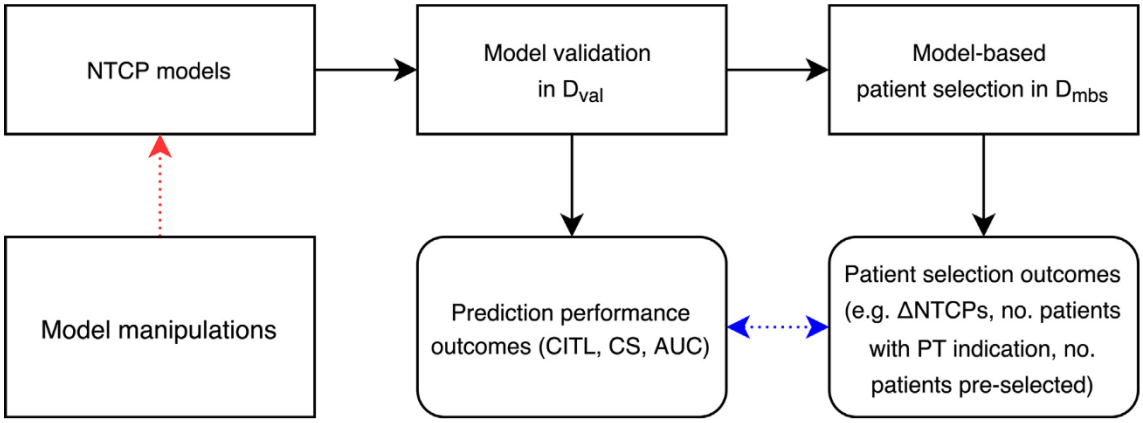


**Fig. 1.** Simulation study flow of the model-based selection process. We manipulate the models such that each prediction performance outcome is independently affected, and relate changes in prediction performance at the validation stage to changes in patient selection. Dval; the validation dataset, Dmbs; the data in which model-based selection is performed, CITL; calibration-in-the-large, CS; Calibration slope, AUC; Area under the receiver operating characteristic curve, NTCP; normal-tissue complication probability, PT; proton therapy, ΔNTCP; difference in risk based on the photon versus the proton plan.

**Table 1**
External validation results.

| Model | Area under the receiver operating characteristic curve (1 is best) | Calibration-in-the-large (0 is best) | Calibration slope (1 is best) |
|---|---|---|---|
| A | 0.71 | 0.0 | 1.0 |
| B | 0.70 | 0.3 | 0.7 |
| C | 0.74 | 0.2 | 1.2 |

model quality (hereinafter $D_{val}$), and one empirical dataset collected at a later point in time in which the model-based patient selection procedure was actually applied (henceforth $D_{mbs}$). $D_{val}$ consists of the data originally used to develop and validate the models in the NIPP: data from HNC patients receiving photon-based radiotherapy (conformal radiotherapy; 3D-CRT, intensity-modulated radiotherapy; IMRT, or volumetric modulated arc therapy; VMAT) of whom 893 from the University Medical Center Groningen (UMCG), 200 from MAASTRO Clinic, and 52 from the Radiotherapeutic Institute Friesland (RIF), between 2007 and 2017. $D_{mbs}$ contains data of 289 patients receiving either VMAT or IMPT (intensity-modulated proton therapy) who underwent a treatment plan comparison at the UMCG between 2018 and 2020. For these 289 patients both their VMAT and the IMPT plans are available and used to calculate the ΔNTCPs per model and retrospectively simulate the MBS procedure under different models and compare the new selection to the treatment originally received. Characteristics of both datasets are presented in Table 3. Further details are provided in Supplementary File A.

*Model-based patient selection*

The MBS procedure in this study uses the models and thresholds defined in the NIPP v2.2 (shown in Fig. 2). Patients receive an indication for PT if the predicted complication risk for their PT-plan is sufficiently lower than the predicted complication risk given their RT-plan, i.e., if their ΔNTCP (i.e., $NTCP_{RT} - NTCP_{PT}$) is sufficiently large, considering the complications and thresholds stated in the NIPP (and Fig. 2.).

**Table 3**
Study data characteristics.

| | $D_{val}$ (n = 1145) | $D_{mbs}$ (n = 289) |
|---|---|---|
| Sex (female) | 295 (25.8 %) | 61 (21.1 %) |
| Mean age (SD) | 63 (10) | 64 (10) |
| Primary tumor location (%) | | |
| Oral cavity | 66 (5.8 %) | 27 (9.3 %) |
| Oropharynx | 411 (35.9 %) | 119 (41.2 %) |
| Nasopharynx | 45 (3.9 %) | 14 (4.8 %) |
| Hypopharynx | 121 (10.6 %) | 27 (9.3 %) |
| Larynx | 502 (43.8 %) | 102 (35.3 %) |
| T-Stage (%) | | |
| Tis-2 | 557 (48.6 %) | 107 (37.0 %) |
| T3-4 | 588 (51.4 %) | 182 (63.0 %) |
| Xerostomia at month 6 after treatment | | |
| Grade 0–1 (none-little) | 619 (54.1 %) | N/A |
| Grade 2 (moderate) | 344 (30.0 %) | N/A |
| Grade 3 (severe) | 182 (15.9 %) | N/A |
| Dysphagia at month 6 after treatment | | |
| Grade 0–1 (regular diet) | 791 (69.1 %) | N/A |
| Grade 2 (soft food) | 166 (14.5 %) | N/A |
| Grade 3–5 (liquids only or tube feeding dependent) | 188 (16.4 %) | N/A |

N/A = not available.

*Model variations and prediction performance*

To influence the calibration in the large (CITL) we gradually increased or decreased the model intercept. This introduced a general over- or under estimation of risk by the model, and consequently resulted in a CITL < 0 and > 0, respectively. This model variation did not impact calibration slope or discrimination.

To influence the calibration slope (CS) we scaled the model's linear predictor by multiplying it by a certain factor. Scaling the linear predictor with a factor value above 1 results in more

**Table 2**
Coefficients of the original NTCP models from the NIPP used in this study. Dmean; mean dose, PCM; pharyngeal constrictor muscle.

| | Xer2+ | Xer3+ | Dys2+ | Dys3+ |
|---|---|---|---|---|
| | Xerostomia (Q41, EORTC H&N35) 6 months after the end of treatment: grade 2 or higher | Xerostomia (Q41, EORTC H&N35) 6 months after the end of treatment: grade 3 or higher | Dysphagia (CTCAE), 6 months after the end of treatment: grade 2 or higher | Dysphagia (CTCAE), 6 months after the end of treatment: grade 3 or higher |
| Intercept ($\beta_0$) | −2,2951 | −3,7286 | −4,0536 | −7,6174 |
| √(Dmean Parotis ipsilateral) + √(Dmean Parotis contralateral) | 0,0996 | 0,0855 | | |
| Dmean of both submandibularis | 0,0182 | 0,0156 | | |
| Dmean oral cavity | | | 0,0300 | 0,0259 |
| Dmean PCM superior | | | 0,0236 | 0,0203 |
| Dmean PCM medius | | | 0,0095 | 0,0303 |
| Dmean PCM inferior | | | 0,0133 | 0,0341 |
| Baseline xerostomia: none | 0,0000 | 0,0000 | | |
| Baseline xerostomia: some | 0,4950 | 0,4249 | | |
| Baseline xerostomia: moderate-severe | 1,2070 | 1,0361 | | |
| Baseline dysphagia: grade 0–1 | | | 0,0000 | 0,0000 |
| Baseline dysphagia: grade 2 | | | 0,9382 | 0,5738 |
| Baseline dysphagia: grade 3–5 | | | 1,2900 | 1,4718 |
| Primary tumor location: oral cavity | | | 0,0000 | 0,0000 |
| Primary tumor location: pharynx | | | −0,6281 | 0,0387 |
| Primary tumor location: larynx | | | −0,7711 | −0,5303 |

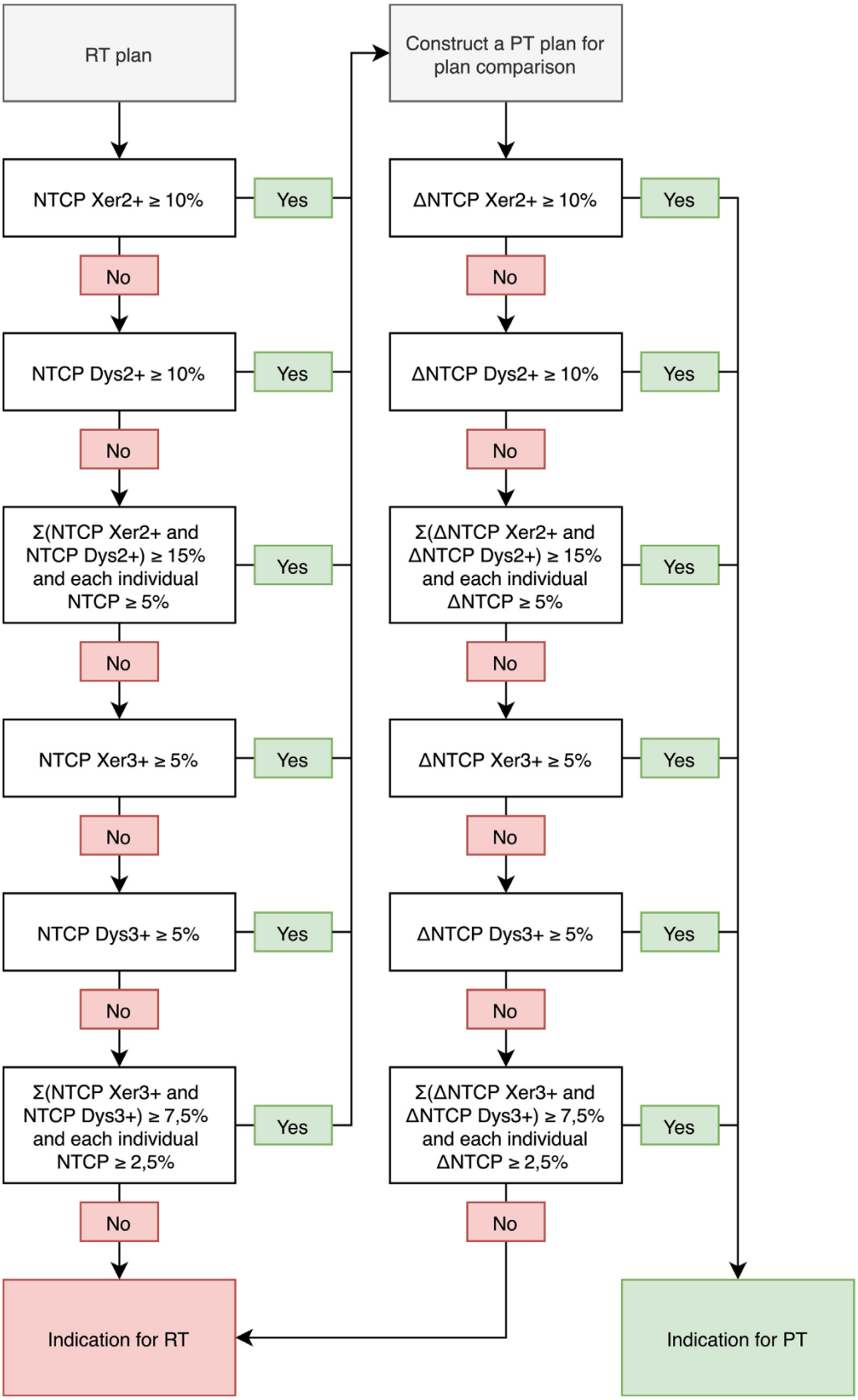


**Fig. 2.** The model-based selection procedure to come to a treatment indication[18]. RT; radiotherapy, PT; proton therapy, NTCP; normal-tissue complication probability, Xer; xerostomia, Dys; dysphagia, 2+; ≥ grade 2, 3+; ≥ grade 3, ΔNTCP; difference in risk based on the photon versus the proton plan.

extreme predictions (closer to 0 or 1) resulting in a CS < 1 (underestimation for low risk patients and overestimation in high risk patients), while a factor value below 1 results in more conservative predictions (closer to the outcome prevalence), resulting in a CS > 1 (overestimation for low risk patients and underestimation for high risk patients). Scaling the linear predictor may also change the CITL, so to make sure we could address the impact of a change in CS only, we additionally recalibrated the model intercept to the predictions on $D_{val}$ of the original model to obtain the original CITL using an offset model after each manipulation of CS.

To influence the area under the receiving operator characteristic curve (AUC) we updated the regression coefficients of the predictors in the model by increasingly interpolating them with uninformative random noise. This affects how well the model ranks patients with regard to their risk, as ranking becomes more arbitrary when more noise is introduced. For each coefficient the noise is sampled from a zero-centered unit-variance normal distribution. As this interpolation may also influence the model calibration, we also recalibrated the model's slope and intercept on the original predictions in $D_{val}$ to obtain the same CS and CITL as the original model.

### *Patient selection outcome measures*

To assess the impact of each model adjustment (and corresponding change in performance measure) on selection, we measured the following statistics on the $D_{mbs}$.

1. The mean ΔLP (i.e., $LP_{RT}$ – $LP_{PT}$, with LP being the linear predictor)
2. The mean predicted NTCP for the RT plans, and for the PT plans.
3. The mean ΔNTCP (i.e., $NTCP_{RT}$ – $NTCP_{PT}$) based on these plans.
4. Changes in ΔNTCP (as root mean squared error), for fixed deviations in prediction performance: −0.2 and 0.2 for CITL and CS and −0.02 for AUC. These values were chosen for them to be likely to be encountered during the validation of a model in practice.
5. The total number of patients receiving RT or PT, based on the adjusted model following the selection procedure in the current NIPP.
6. The number of patients switching from RT to PT and vice versa, based on the adjusted model following the selection procedure in the current NIPP.

## Results

### *Assessment of model variations*

We confirmed that the model adjustments did influence each prediction performance measure independently of the others (shown in Supplementary file B), meaning we were able to assess the *individual* impact of each of the performance measures on the patient selection outcomes.

### *Impact on mean ΔNTCP*

In the first column of Fig. 4, across models, increasing CITL resulted in a lower mean $NTCP_{RT}$ and mean $NTCP_{PT}$, and no change in mean ΔLP. This is by construction, as CITL directly reflects the mean of predicted risks relative to the mean proportion of observed outcomes. An increase in CITL results for models Xer3+, Dys2+ and Dys3+ in a steady decrease in mean ΔNTCP. For Xer2+ the mean ΔNTCP is almost flat, with a slight upward bend around NTCPs of 50 %. Given that only the model intercept has changed in the CITL variation, and thus the ΔLPs remain constant, the changes in mean ΔNTCP are explained by the shift in NTCP (as illustrated in Explanation Box 2). The ΔNTCP results follow the first derivative shown in Fig. 3 (the red dashed line), showing a slight increase in mean ΔNTCP around a mean NTCP of 50 % (for Xer2+), but a steady decrease in mean ΔNTCP as mean NTCP approaches 0 (particularly for Xer3+, Dys2+ and Dys3+). Lastly, in general changes in CITL are larger for the dysphagia models because they have larger ΔLP's to begin with.

For all models, with an increase in CS (column 2 in Fig. 4) we observed a decrease in mean ΔLP and an increase in mean $NTCP_{RT}$

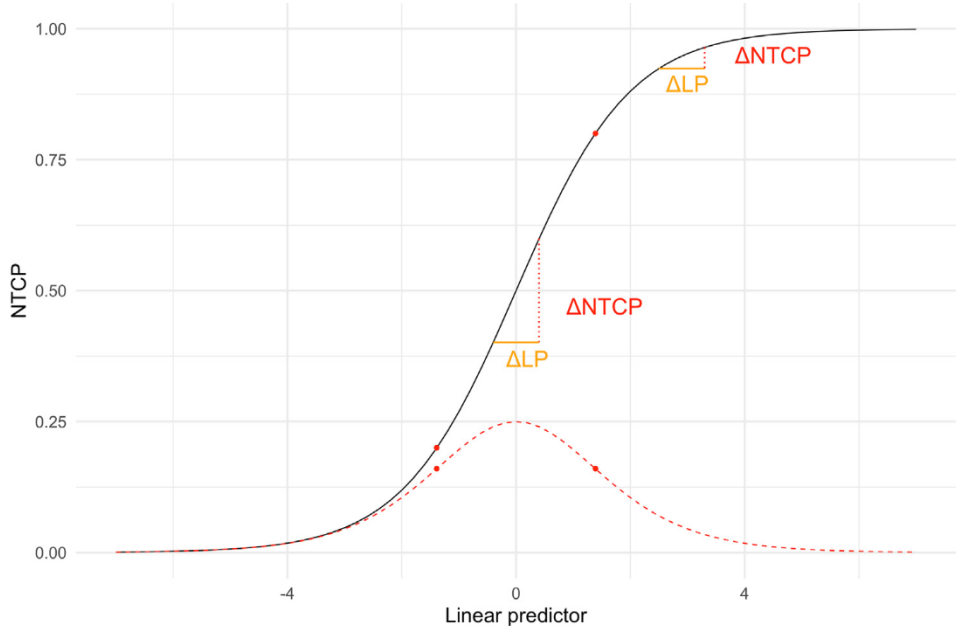


**Fig. 3.** The logistic curve (black solid line) and its first derivative (the dashed red line).

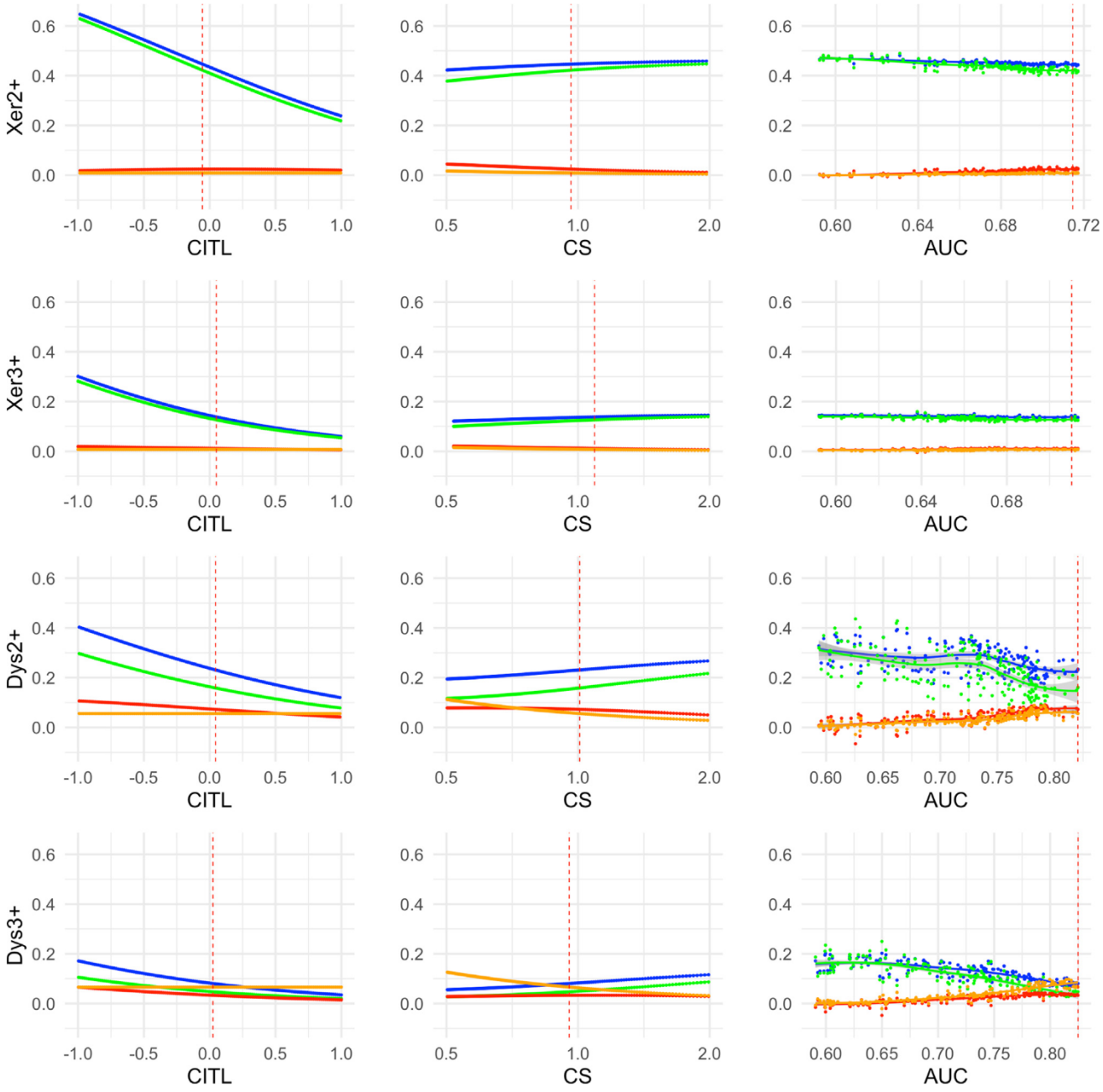


**Fig. 4.** For each of the four predicted outcomes (one per row), we plotted the mean $NTCP_{RT}$ (blue), the mean $NTCP_{PT}$ (green), the mean $\Delta LP \cdot 10^{-1}$ (orange; $\cdot 10^{-1}$ is purely so $\Delta LP$ can be shown in the same plot), and the mean $\Delta NTCP$ (red) in $D_{mbs}$ against the model prediction performance in $D_{val}$ (on the x-axis), in terms of CITL, CS or AUC. The vertical dashed red line in each plot indicates the performance of the original model in $D_{val}$.

and $NTCP_{PT}$. For all models but Dys3+, we observed that an increase in CS (column 2 in Fig. 4) resulted in a decrease in mean $\Delta NTCP$ (decrease in the red line). Important to notice when interpreting these results is that - in isolation - a decrease in $\Delta LP$ would result in lower $\Delta NTCP$s, while NTCPs getting closer to 50 % would increase $\Delta NTCP$ (see Explanation Box 2). These two may balance each other out resulting in relatively small changes in mean $\Delta NTCP$. Later in the article we also discuss patient-level changes in $\Delta NTCP$ (instead of changes in the mean $\Delta NTCP$).

A decrease in AUC resulted in a decrease in $\Delta NTCP$ across settings, converging to a mean $\Delta NTCP$ of 0 as the AUC approaches 0.5 (no meaningful discrimination between patient risks). For all settings this can be attributed to the decrease in $\Delta LP$.

*Patient-level changes*

In Fig. 5, the $\Delta NTCP$s of the original model (on the y-axis) are plotted against the $\Delta NTCP$ of a manipulated model (on the x-axis) for deviations in CITL and CS of −0.2 and 0.2, and for a deviation in AUC of −0.02. The $\Delta NTCP$ treatment indication thresholds for PT are indicated as red lines.

*Changes in ΔNTCP*

First, we observed the ranking of the performance measures per column (as RMSE cannot be compared between outcomes with different initial $\Delta NTCP$ distributions). Across outcomes a deviation in AUC led to the largest RMSE (last row), followed by a decreased CS (third row). Particularly for AUC this can also be observed visually.

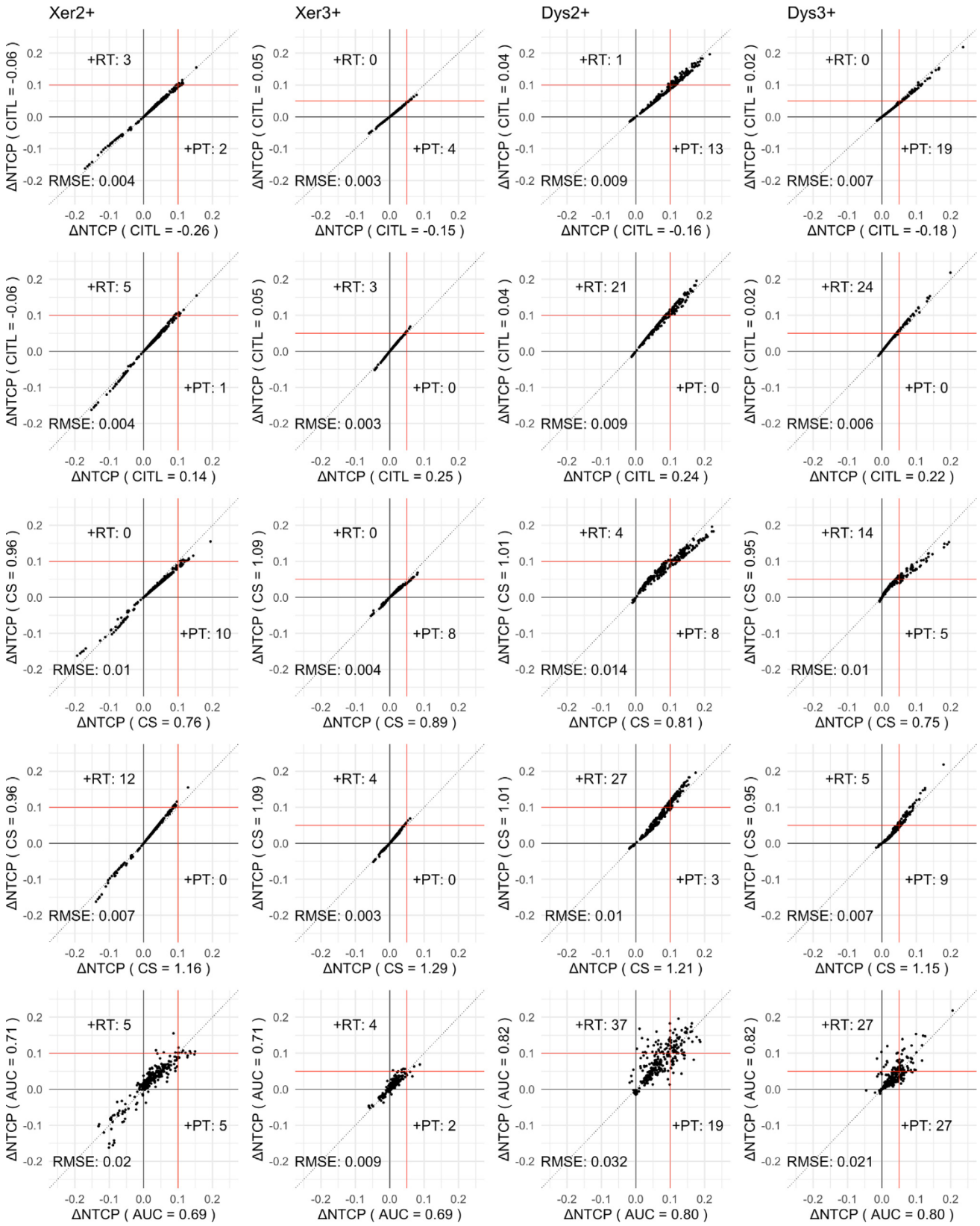


**Fig. 5.** Per patient in $D_{mbs}$, for each predicted outcome and type of manipulation, the ΔNTCPs of the original model (y-axis) were plotted against the ΔNTCP of a manipulated model (x-axis). Root mean squared error (RMSE) between the two ΔNTCPs has been calculated (lower RMSE indicates higher agreement). Red lines indicate the ΔNTCP thresholds from the NIPP. +RT and +PT indicate the number of patients that would receive a different indication by the manipulated model compared to the original - if *only* that model was used to determine treatment indication (NB: in practice, the four models are used conjointly for patient selection).

The ranking regarding deviations in CITL or an increased CS were less consistent. The direction of change in ΔNTCP was consistent with those discussed at Fig. 4: a decrease in CITL resulted in higher ΔNTCP (in the top row most points lie below the diagonal), while an increase in CITL resulted in lower ΔNTCP (in the second row most point lie above the diagonal). Although overall an increase in CS or decrease in AUC were observed to lead to a decrease in mean ΔNTCP (in Fig. 4), on a patient level, deviations in CS or AUC resulted in both increased ΔNTCP as well as decreased ΔNTCP for certain patients. This is because for deviations in CS and AUC per patient the ΔLP as well as the absolute LPs (or NTCPs) may change. This is in contrast to deviations in CITL, where changes in ΔNTCP only relate to changes in LP (or NTCP), as the ΔLP does not change.

*Changes in treatment indication*

For each manipulation we observed whether they could lead to changes in treatment indication if only that model was used (in practice all four models are used as shown in Fig. 2). The changes in treatment indications are consistent with the changes in mean ΔNTCP (from Fig. 4). Whether a change in ΔNTCP leads to a change in treatment indication depends on the ΔNTCP thresholds. A deviation of −0.02 in AUC led to relatively large changes in treatment indication across outcomes. The direction of change was mixed: for Xer3+, Dys2+ and Dys3+ more patients were indicated with RT, while for Xer2+ more patients were assigned PT.

When inspecting changes in treatment indication including potential interplay between the four models, we observe that the overall effects follow expectations based on mean changes in ΔNTCP (from Fig. 4). We observe that in general, changes in treatment indications are largest under manipulation of the Dys2+ model, followed by Xer2+, while changes in performance of Dys3+ and Xer3+ had less impact, possibly because most patients passing the thresholds for Dys3+ and Xer3+ may already have indicated PT based on their ΔNTCP for Dys2+ or Xer2+. Further details about this analysis are shown in Supplementary File B.

## Discussion

We investigated how recommended prediction performance measures (AUC, CITL, and CS) relate to changes in ΔNTCP and consequent patient indications in a model-based treatment indication setting. We found that all studied performance measures independently impacted both ΔNTCP and consequent treatment indications. Manipulations of AUC resulted in quite large changes in individual ΔNTCPs. If predicted NTCPs are close to 50 %, CITL had a relatively small impact on ΔNTCP. Changes in AUC and CS may increase or decrease ΔNTCP: some patients receive higher ΔNTCP and others lower ΔNTCP. The direction of change in ΔNTCP due to a change in CITL can be directly read from the slope of the first derivative of the logistic curve (illustrated in Explanation Box 2). Whether changes in ΔNTCP lead to changes in treatment indications depends on: (1) the used thresholds, and (2) the importance of that model in relation to the other used models in the complete treatment indication thresholding scheme. For the studied version of the NIPP (v2.2), manipulation of the Dys2+ model had the largest impact on treatment indications, followed by Xer2+, Dys3+ and Xer3+.

Box 2 Mechanisms behind changes in ΔNTCP. The ΔNTCP depends on (1) the ΔLP, i.e., the difference in linear predictor between the RT and PT plan, reflecting both Δdose and coefficient size, and (2) the absolute RT and PT-based NTCPs. If either through a change in Δdose or a change in model coefficients the ΔLP changes, so does the ΔNTCP (and possibly the treatment indication). The second mechanism of how ΔNTCP may change is because in a logistic NTCP model the same ΔLP does not always result in the same ΔNTCP. This can be observed from the first derivative of the logistic curve (the red dashed line in Fig. 3). For a fixed ΔLP, the ΔNTCP is largest for NTCPs around 0.50, at the top of the first derivative. When NTCPs lie further from 0.50, the same ΔLP will result in lower ΔNTCP.

Moreover, the *change* in ΔNTCP due to a change in NTCP can be linked to the *slope* of the first derivative of the logistic curve, which is steepest around NTCPs of 0.20 and 0.80 (indicated by the red dots in Fig. 3). In other words, for a given ΔLP, if NTCPs are around 0.20 or 0.80, a small change in that NTCP (e.g., due to model miscalibration) results in a relatively large change in ΔNTCP.

To our best knowledge, the current study is the first to relate differences in performance measures to predicted risk differences (such as ΔNTCP) and thus treatment indication decisions. Extensive guidance exists for the validation of prediction models in general[21,23–25], and for NTCP models in particular[26]. However, these articles do not cover the evaluation with regard to risk differences. Hoogland et al.[27] addressed the use of prediction models for individualized treatment effect prediction via risk differences comparing various modeling approaches in a full simulation study. Their study focused on the direct evaluation of treatment effect predictions and not on measures of patient selection. In their conclusions, they explicitly pointed out the need for further work on procedures to evaluate prediction models for individualized treatment decisions. Hayward et al.[28] studied detection of treatment effect heterogeneity via prediction models as an alternative to subgroup analysis in clinical trials and found that models with AUC < 0.6 did not achieve even moderate statistical power (30 % to 45 %) to detect heterogeneity. They therefore recommend using models with AUC > 0.6. They did not study the impact of calibration on patient selection for treatments nor the prioritization of performance measures.

To appreciate the current study results several issues should be addressed. First, the impact of performance measures on changes in ΔNTCP is in part related to the used link function of the models (here: logistic). Our choice for the logistic link function is motivated by the fact that the majority of currently existing NTCP models are based on these functions.[1] While we conjecture that the conclusions may generalize to a broader class of methods that use a logistic function to obtain probabilities (e.g., neural networks with logistic final activations) this requires further research.

Second, all prediction performance measures were studied in isolation, to clearly separate their impact on ΔNTCP. However, in practice one is likely to encounter interacting combinations of imperfect performance measures. Future research is still needed in this regard. Third, while the impact of the performance measures on changes in ΔNTCP is primarily related to the link function of the models and may generalize to other settings, the relation

between ΔNTCP and consequent treatment decisions depends on the used indication protocol, which can differ across centers or countries.[16] Finally, important to note is that we studied the impact of deteriorating performance measures on *changes* in ΔNTCP and also on treatment indication or patient selection, but how the changes affected overall treatment quality remains unclear, and may involve a broad spectrum of toxicity outcomes.[29]

## Conclusions

Since all studied performance measures (AUC, CS and CITL) had clear independent impact on both ΔNTCP and on treatment indications we conclude they should all be assessed thoroughly during model validation for model-based patient selection. While we found the AUC to have quite large impact on changes in ΔNTCP and treatment indication, in view of the relevant literature[26], one should mind that: (1) the upper bound of the AUC is defined by the information in the predictor variables included in the model, and so a perfect AUC can not be realistically expected, (2) AUC depends on cohort composition and can thus only be compared across models when it is calculated for each model in a similar cohort, and (3) the validation cohort should be sufficiently large[30], as confidence intervals of AUC narrow slowly when increasing the validation sample size[26]. Calibration measures may be easier to control in practice, e.g., by recalibration.[31,32] As a practical approach to the selection of models used for patient selection we suggest to perform external validation of multiple models predicting the same outcome in a sufficiently large[30] sample representative of the target patient population and setting, and then consider the model with the highest AUC, and recalibrate it if needed in those data. In the process of arriving at high quality NTCP models for determining model-based treatment indications, besides considering the importance of each performance measure on ΔNTCP, their interaction with treatment thresholds also plays an important role. And, in case multiple models are used to determine treatment indication, the relative importance of each model in the decision process is relevant as well, and can provide prioritization regarding their validation.

## Ethics approval

As the Dutch Medical Research Involving Human Subjects Act is not applicable to data collection as part of routine clinical practice, the requirement of informed consent was waived by the ethics committee.

## Author contributions

AL, JR, KM, and ES, conceived the study. AL designed and carried out the simulation studies, and drafted the first version of the manuscript. JL, LB, AS, FH, and OW provided the study data. All authors contributed to the writing and approval of the final version.

## Data availability

The computer code used to conduct the experiments is available at https://github.com/tuur/NTCPPPmeasures. The original patient data is not available for privacy reasons.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: J.A. Langendijk reports a relationship with Dutch Cancer Society: funding grants. J.A. Langendijk reports his department has research contracts with IBA, RaySearch, Siemens, Elekta, Leoni, and Mirada. J.A. Langendijk reports a relationship with Global Scientific Advisory Board of IBA, RayCare International Advisory Board of RaySearch that includes: board membership, consulting or advisory, and speaking and lecture fees. J.A. Langendijk reports a relationship with Netherlands Society for Radiation Oncology that includes: board membership. J.B. Reitsma and E. Schuit are involved as methodologist in the development of indication protocols for patient selection for proton therapy in the Netherlands.

## Acknowledgements

The HTx project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 825162. This dissemination reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

## Conflicts of Interest

J.B. Reitsma: Involvement as methodologist in the development of indication protocols for patient selection for proton therapy in the Netherlands.

J.A. Langendijk: Department has research contracts with IBA, RaySearch, Siemens, Elekta, Leoni, and Mirada. Received grants from Dutch Cancer Society and EU. Member of Global Scientific Advisory Board of IBA. Member of RayCare International Advisory Board of RaySearch. Chair of the Netherlands Society for Radiation Oncology.

E. Schuit: Involvement as methodologist in the development of indication protocols for patient selection for proton therapy in the Netherlands.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2022.109449.

## References

[1] Sharabiani M, Clementel E, Andratschke N, Hurkmans C. Generalizability assessment of head and neck cancer NTCP models based on the TRIPOD criteria. Radiother Oncol 2020;146:143–50.
[2] Brodin NP, Kabarriti R, Garg MK, Guha C, Tomé WA. Systematic review of normal tissue complication models relevant to standard fractionation radiation therapy of the head and neck region published after the QUANTEC reports. Int J Radiat Oncol 2018;100:391–407.
[3] Rodrigues G, Lock M, D'Souza D, Yu E, Van Dyk J. Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review. Radiother Oncol 2004;71:127–38.
[4] Stieb S et al. NTCP modeling of late effects for head and neck cancer: a systematic review. Int J Part Ther 2021;8:95–107.
[5] Dawson LA et al. Analysis of radiation-induced liver disease using the Lyman NTCP model. Int J Radiat Oncol 2002;53:810–21.
[6] Takada T et al. Prognostic models for radiation-induced complications after radiotherapy in head and neck cancer patients. Cochrane Database Syst Rev 2021. https://doi.org/10.1002/14651858.CD014745.
[7] Witte MG et al. IMRT optimization including random and systematic geometric errors based on the expectation of TCP and NTCP. Med Phys 2007;34:3544–55.
[8] Kierkels RGJ et al. Multivariable normal tissue complication probability model-based treatment plan optimization for grade 2–4 dysphagia and tube feeding dependence in head and neck radiotherapy. Radiother Oncol 2016;121:374–80.
[9] Zaider M, Amols HI. Practical considerations in using calculated healthy-tissue complication probabilities for treatment-plan optimization. Int J Radiat Oncol 1999;44:439–47.
[10] Christianen MEMC et al. Swallowing sparing intensity modulated radiotherapy (SW-IMRT) in head and neck cancer: clinical validation according to the model-based approach. Radiother Oncol 2016;118:298–303.

[11] van der Laan HP, Christianen MEMC, Bijl HP, Schilstra C, Langendijk JA. The potential benefit of swallowing sparing intensity modulated radiotherapy to reduce swallowing dysfunction: an in silico planning comparative study. Radiother Oncol 2012;103:76–81.
[12] Marks LB et al. Use of normal tissue complication probability models in the clinic. Int J Radiat Oncol 2010;76:S10–9.
[13] Langendijk JA et al. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol 2013;107:267–73.
[14] Dutz A et al. Identification of patient benefit from proton beam therapy in brain tumour patients based on dosimetric and NTCP analyses. Radiother Oncol 2021;160:69–77.
[15] Zientara N, Giles E, Le H, Short M. A scoping review of patient selection methods for proton therapy. J Med Radiat Sci 2022;69:108–21.
[16] Tambas M et al. Current practice in proton therapy delivery in adult cancer patients across Europe. Radiother Oncol 2022;167:7–13.
[17] Langendijk JA et al. National protocol for model-based selection for proton therapy in head and neck cancer. Int J Part Ther 2021;8:354–65.
[18] Landelijk Platform voor Radiotherapie bij Longtumoren & Landelijk Platform Protonentherapie. Landelijk Indicatie Protocol Protonentherapie Longcarcinoom. (2019).
[19] Tambas M et al. First experience with model-based selection of head and neck cancer patients for proton therapy. Radiother Oncol 2020;151:206–13.
[20] Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. JAMA Psychiat 2020;77:534–40.
[21] Steyerberg EW et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381.
[22] den Bosch LV et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. Radiother Oncol 2020;148:151–6.
[23] Van Calster B et al. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230.
[24] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.
[25] Moons KGM et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–W.
[26] Bahn E, Alber M. On the limitations of the area under the ROC curve for NTCP modelling. Radiother Oncol 2020;144:148–51.
[27] Hoogland J et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Stat Med 2021;40:5961–81.
[28] Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Method 2006;6:18.
[29] Van den Bosch L et al. Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: a new concept for individually optimised treatment. Radiother Oncol 2021;157:147–54.
[30] Riley RD et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 2021;40:4230–51.
[31] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008;61:76–86.
[32] Vergouwe Y et al. A closed testing procedure to select an appropriate method for updating prediction models. Stat Med 2017;36:4529–39.