Patient selection for proton therapy using NTCP estimation after fast dose prediction with deep learning for head and neck cancer

Margerie Huet-Dastarac[1], Steven Michiels[1], Sara Teruel Rivas[1], Hamdiye Ozan[1], Edmond Sterpin[1,2], John A. Lee[1] ,Ana Barragan-Montero[1]

[1]Molecular Imaging, Radiotherapy, and Oncology (MIRO), IREC, UCLouvain, Brussels, Belgium

[2]Department of Oncology, Laboratory of Experimental Radiotherapy, KU Leuven, Leuven, Belgium.

*** Correspondence:**

Corresponding Author

margerie.huet@uclouvain.be

Keywords: Deep Learning, Proton Therapy, Radiotherapy, Radiation effects, Adverse effects. Patient Selection or NTCP.

Abstract

Background

In cancer care, determining the most beneficial treatment technique is a key decision affecting the patient's survival and quality of life. Patient selection for proton therapy (PT) over conventional radiotherapy (XT) currently entails comparing manually generated treatment plans, which requires time and expertise.

Purpose

We developed an automatic and fast tool, AI-PROTIPP (Artificial Intelligence Predictive Radiation Oncology Treatment Indication to Photons/Protons), that assesses quantitatively the benefits of each therapeutic option. Our method uses deep learning (DL) models to directly predict the dose distributions for a given patient for both XT and PT. By using models that estimate the Normal Tissue Complication Probability (NTCP), namely the likelihood of side effects to occur for a specific patient, AI-PROTIPP can propose a treatment selection quickly and automatically.

Material and Methods

A database of 60 patients presenting oropharyngeal cancer, obtained from the Cliniques Universitaires Saint Luc in Belgium, was used in this study. For every patient, a PT plan and an XT plan were generated. The dose distributions were used to train the two dose DL

prediction models (one for each modality). The model is based on U-Net architecture, a type of convolutional neural network currently considered as the state of the art for dose prediction models. A NTCP protocol used in the Dutch model-based approach, including grade II and III xerostomia and grade II and III dysphagia, was later applied in order to perform automatic treatment selection for each patient. The networks were trained using a nested cross-validation approach with 11 folds. We set aside 3 patients in an outer set and each fold consists of 47 patients in training, 5 in validation and 5 for testing. This method allowed us to assess our method on 55 patients (5 patients per test times the number of folds).

Results

The treatment selection based on the DL-predicted doses reached an accuracy of 87.4% for the threshold parameters set by the Health Council of the Netherlands. The selected treatment is directly linked with these threshold parameters as they express the minimal gain brought by the PT treatment for a patient to be indicated to PT. To validate the performance of AI-PROTIPP in other conditions, we modulated these thresholds and the accuracy was above 81% for all the considered cases. The difference in average cumulative NTCP per patient of predicted and clinical dose distributions is very similar (less than 1% difference).

Conclusion

AI-PROTIPP shows that using DL dose prediction in combination with NTCP models to select PT for patients is feasible and can help to save time by avoiding the generation of treatment plans only used for the comparison. Moreover, DL models are transferable, allowing experience to be shared with centers that would not have PT planning expertise.

1 Introduction

In cancer care, determining the most beneficial treatment technique is a key decision affecting the patient's survival and quality of life. Having an automatic and fast tool to assess quantitatively the benefits of each therapeutic option would lead to an improved treatment outcome and better use of medical resources. In radiation therapy, the choice between proton therapy (PT) or conventional radiotherapy (XT) is a glaring example, as PT has the potential of reducing toxicity in healthy tissues with an equivalent target coverage, but at a higher cost and reduced availability (Brodin et al. 2021). The so-called "model-based" approach, developed in the Netherlands, has become a popular solution for treatment indication. Patients are directed towards PT or XT based on models of Normal Tissue Complication Probability (NTCP) (Langendijk et al. 2013; Landelijk Indicatieprotocol Protonent...). These NTCP models enable the radiotherapist in charge to estimate probabilities of observing side effects and therefore compare XT and PT plans quantitatively (Tambas et al. 2020).

The model-based approach typically requires the generation of at least one plan for each of the modalities under comparison, namely, XT and PT, to subsequently feed the NTCP models with the corresponding dose distributions. However, the generation of a treatment plan is a well known bottleneck when it comes to accelerating and standardizing the workflow. Treatment planning involves numerous manual operations and its duration is counted in hours or even days. Long generation time is not the only shortcoming of the current planning process: variability in the planner's skill level may lead to reduced plan quality (Nelms et al. 2012), compromising the

treatment indication. This effect might be even more pronounced in PT planning, since the number of centers delivering PT is well inferior to XT centers. According to (Abdel-Wahab et al. 2021) and (Kazantsev et al. 2021), there were more than 7000 radiation therapy centers in 2021 and less than 100 were delivering particle therapy treatments. Hence, the expertise needed to generate a PT plan for a specific location is not easily accessible for new PT centers or centers without PT, which often obstructs the patient selection process.

This work addresses this cost in time and resources during the treatment plan optimization process, the accessibility problems to PT expertise and the risk of operator variability in the decision between PT or XT plans. We apply to oropharyngeal cancer, a type of head and neck (H&N) cancer, as a proof-of concept that can be translated to other treatment locations provided that we have a database of plans and a well established NTCP protocol. We focus on H&N as it is one of the most complex locations to plan and oropharyngeal cancer is the most frequent H&N cancer (Leroy et al. 2019). For that purpose, we take advantage of deep learning (DL) models to directly predict the optimal XT and PT dose distributions for a given patient, and later feed NTCP models to perform the final treatment choice. These DL dose prediction models are fast (i.e., they predict the dose in a few seconds instead of hours to obtain plans manually), they can be shared easily, and they have shown to achieve a high accuracy for different cancer locations [6–9]. In addition, if properly trained with high quality databases (Barragán-Montero et al. 2021), DL models are an excellent tool to capture the planning expertise from centers with accumulated clinical experience and transfer that knowledge to new or less experienced centers. This mitigates the problem of operator variability and ensures a homogeneous clinical practice.

Although the literature about DL dose prediction is rich, the application of these models for treatment selection is scarce. A recent study has explored the use of dose prediction models to choose between XT and PT treatment for pediatric abdominal tumors (Guerreiro et al. 2021), but the combination of DL dose prediction with NTCP models remains to be addressed. In particular, for H&N cancer, where NTCP models have been well validated (Tambas et al. 2020), their combination with DL dose prediction models has the potential to significantly improve the treatment selection process. Two studies tackle H&N for proton treatment selection. One using a predictive tool for the mean dose (Dmean) of organs at risk (OAR) depending on the distance to the target volume (Tambas et al. 2022). The other one uses the RapidPlan system with a regression analysis generated from former clinically approved treatment plans (Hytönen et al. 2022). RapidPlan uses such a regression algorithm based on dose-volume histogram (DVH) estimation. The DVH estimate collapses 3D information in 2D metrics (dose and volume), losing the information about its spatial distribution (Ahn et al. 2021). In ICRU 83, the guidelines state that due to this limitation, a slice-by-slice inspection of the dose distribution is recommended to identify potential aspects for further improvement (Hodapp 2012). Our proposed method performs treatment selection based on DL dose prediction, which yields a 3D dose distribution of the selected plan and can be integrated in a fully automated AI-based workflow. We named it AI-PROTIPP, which stands for AI-Predictive Radiation Oncology Treatment Indication to Photons/Protons.

2 Materials and Methods

2.1 Data and prior dose distribution

An anonymized database of 60 patients with oropharyngeal cancer is used to train the DL dose prediction networks. The data was obtained from Cliniques Universitaires of Saint Luc (Brussels, Belgium). For each patient, a PT plan and XT plan were manually generated by an experienced dosimetrist. Local ethics committee approval has been obtained for the collection and use of this retrospective database.

All treatments are planned in the treatment planning system (TPS) RayStation 11B (Raysearch Laboratories, AB) with a research license. For consistency regarding the planning strategy and the beam configuration, all selected patients are bilateral cases and they have two dose prescription levels: 70Gy (TV_HIGH) and 54.25Gy (TV_LOW) on tumor and elective nodal volumes, respectively.

PT plans are planned for pencil beam scanning delivery, using four beams at (10º, 60º), (10º, 120º), (350º, 240º) and (350º, 300º) for couch and gantry angles, respectively. The PT plans are generated using Monte Carlo dose calculation and robust worst-case optimization (Souris et al. 2019) on the CTV volume with 21 scenarios (2.6% proton range error and 4mm systematic error in the three space directions).

The XT plans are planned for volumetric modulated arc therapy (VMAT). The CTV to PTV margin equaled 4mm.

Each patient has a CT image and a set of contours for the target volumes (CTV or PTV) and the OARs: brainstem, spinal cord, parotids, esophagus, glottic area, oral cavity, pharyngeal constrictor muscles inferior, medium and superior, submandibular glands and supraglottic larynx. The last input of the models is a prior knowledge of the dose distribution, which is obtained by running a first automatic optimization using a template of the objective functions in the TPS without any manual tuning. This first estimate of the dose distribution follows a template of dose volume objectives for the tumor site. In a way, the network is allowed to focus on the task of the dosimetrist: find the optimal dose starting from this estimation.

The described patient data was exported from the RayStation software in DICOM format and converted into NumPy arrays to feed our DL models that are trained on our servers.

2.2 Network architecture and training

The DL model used for dose prediction is a HD-Unet (Nguyen et al. 2019), a type of convolutional neural network, whose architecture is a hybrid version between the popular U-Net (Ronneberger et al. 2015) and DenseNet (Huang et al. 2017). U-Net is an architecture which has shown outstanding performance in image-to-image mapping and is therefore very popular. It was first published in 2015 and since then has become the state of the art for dose prediction (Liu et al. 2020), (Barragán-Montero et al. 2019). U-Net architectures can include global and local features to generate 3D voxel-wise predictions by working at different resolution levels, while dense connections (a feature of DenseNets) improve the learning by enhancing feature reuse as well as gradient backpropagation. Due to the increased memory usage of dense connections, the HD-Unet only uses them within levels of the same resolution. Two HD U-Nets are trained separately. One with the PT database and one with the XT. Both models are fed with input data that gather the prior dose distribution, the CT scan, the prescriptions of the target volumes and the binary matrices containing the volume of the OARs. The workflow is shown in (Figure 1).

As illustrated in (Figure 2), we assess the performance of AI-PROTIPP with an 11-fold nested cross-validation approach. This method entails to set aside 3 patients as an outer test set and train multiple times the networks with moving training, validation, and test sets. The validation and test sets consist of 5 patients each, and we use the remaining 47 patients for the training. The process is repeated 11 times, allowing us to obtain predictions on 55 patients (number of patients per test set times number of folds). Since the 3 left out patients are not involved during training or used for determining the network hyperparameters, they serve as an independent test set to evaluate the method. In (Figure 3), we display the variation of the DVH metrics of interest for the 11 predicted doses for the 3 left out patients. For statistical analysis, the predicted doses and planned doses are compared with the paired-samples Wilcoxon test. The threshold for statistical significance is set at p-value < 0.05. All these operations were performed using a Tesla A100 GPU.

2.3 NTCP models

The patient selection protocol used in this study is the last version of the Landelijk Indicatie Protocol Protontherapy (LIPPv2.2), developed for H&N cancer (Landelijk Indicatieprotocol Protonent...). The NTCP models are applied to both the planned and predicted dose distributions for the PT and XT modalities. LIPPv2.2 estimates the probability of occurrence for four distinct late side effects: dysphagia (grade II and III) and xerostomia (grade II and III). Xerostomia probabilities are estimated from the tumor location and the DVH metrics in the parotid and submandibular glands. Dysphagia, on the other hand, involves the dose in the oral cavity and the pharyngeal constrictor muscles (inferior, middle, and superior), as well as the tumor location among : oral cavity, pharynx, and larynx. As for any NTCP model developed in the Netherlands, patients are selected for PT when the estimated NTCP reduction is equal or larger than 10% for complications of grade II and 5% for complications of grade III (Beetz et al. 2012; Christianen et al. 2012). Two other criteria combining grade II side-effects and grade III side-effects are set, respectively. The criteria and respective thresholds are reported in (Figure 1) and if at least one of them is passed, the patient is selected for PT. The NTCP values were computed offline, on our own servers, by extracting the dose metrics needed for the equations in LIPPv2.2. RayStation does not provide NTCP calculation tools but it can be scripted so that there would be a seamless workflow.

2.4 Evaluation of AI-PROTIPP

Equation 1. $weighted\_accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$

where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively. We consider a case positive when the patient is selected for PT. A false positive (FP) would therefore be a patient selected for PT considering the predicted dose but XT with the clinical dose.

We evaluate the performance of AI-PROTIPP like an imbalanced classification problem, by computing the weighted accuracy, Equation 1, of the patient selection when using the default threshold values in the protocol. The advantage of weighted accuracy is that it is more stringent and it will give more importance to the misclassified patients of the underrepresented class.

Contrary to all other parameters of the NTCP models, these thresholds (i.e. 5% and 10%) were not determined by clinical evidence; instead, they are fixed by the Health Council of the Netherlands. The partition of patients selected for PT and XT is directly linked with these threshold parameters

as they express the minimal gain brought by the PT treatment for a patient to be indicated to PT. To validate the performance of our tool in other conditions, we modulate these thresholds which yields different repartitions of patients.

In order to measure the overall impact of the proposed method on our patient cohort, we compute the average cumulative NTCP for patients following the treatment selection from the planned and predicted doses, respectively. The NTCP values for the 4 side effects of every patient are summed separately and averaged by dividing with the number of patients. In other words, if PT is indicated for a patient by the planned doses but XT is indicated by the predicted one, we add the planned PT NTCP to the cumulative value with planned doses and the predicted XT NTCP to the cumulative value with predicted doses. The difference between the final average cumulative values can serve as a measure of the total gain or loss in NTCP over the patient population when using DL-predicted doses instead of manual planning.

3 Results

The DL models predict the dose with a median absolute error below 1.6% in the D95 and D99 of the target volume for XT and PT (Figure 4, bottom). Values are expressed in percentage of the highest prescription dose. The DVH metrics for most OARs, including those involved in the NTCP models, have an interquartile range (IQR) of error below 5% and 3% for XT and PT, respectively (Figure 4, top). The Dmean for OARs involved in the NTCP protocol does not differ significantly ($p<0.05$) between planned and predicted results, neither for XT nor PT (Table 1). Training time is around 6 hours for each network and it takes about 11 seconds to predict the dose distribution of one modality for a patient. We assess the performances of these 11 networks on the 3 unseen patients and the standard deviations of the DVH metrics are less than 3% of the prescription for the three patients.

Regarding the selection process based on NTCP criteria, we first determined the treatment indication based on the clinically planned doses. With the thresholds fixed by the Health Council of the Netherlands, we obtain 85% of patients selected for PT, which is an uneven partition. These results are in line with the clinical study (Yasuda et al. 2021) and are due to the location of the target in the oropharynx. In that context, AI-PROTIPP yielded a weighted accuracy of 87.4% with 41 TP, 7 TN, 1 FP and 6 FN (Figure 6). Out of the 6 different criteria, the third and the fourth (the ones associated with dysphagia) were the only ones responsible for PT selection (Figure 5).

Since only two thresholds were consistently involved in the treatment selection, we decided to modulate them in order to simulate different situations and assess our tool performance. The treatment selection partitions according to the threshold values can be seen in (Figure 7). On the same graph, we can read the weighted accuracy of the treatment selection based on the predicted doses computed with Equation 1. The weighted accuracy is above 81 percent for all considered threshold values. The average cumulative NTCP for each side effect that is considered when using predicted doses is about 1% lower for all considered thresholds compared to average cumulative NTCP using clinical doses (Figure 8).

4 Discussion

In this study, we investigate a novel method, AI-PROTIPP, for treatment selection for H&N cancer patients using DL dose prediction models in the context of the Dutch model-based approach. For both treatment modalities (XT and PT), our prediction models have shown to be very accurate (Figure 3) and can therefore be good candidates to estimate NTCP and to propose a treatment selection.

Patients suffering from oropharyngeal cancers with large target volume and lymph nodes selected on both sides are glaring examples of cases that can highly benefit from the high conformality of PT plans. This is easily understandable as there are numerous OARs, such as the three PCM or oral cavity, that are located in the central area and cannot be spared with XT plans. With the default thresholds set by the Health Council of the Netherlands, 85% of the clinical plans were selected for a PT treatment indeed. For this configuration, AI-PROTIPP achieves a weighted accuracy of 87.4%. We checked, however, if it was not biased by the imbalanced partition of the treatment indication with clinical doses. As mentioned in Section 2.3, if any of the 6 criteria is met, then the considered patient is selected for PT. In (Figure 5), we can see that only criteria 3 and 4, namely, probability estimates for dysphagia of grades II and III to occur, are responsible for PT treatment selection. If these two thresholds values are increased, we observe as expected that the clinical plans partition varies accordingly. (Figure 7) shows that AI-PROTIPP performs well for different preset partitions of the clinical treatment indication as the weighted accuracy stays above 81%.

In addition, the difference in average cumulative NTCP when considering clinical dose and predicted doses for the treatment indication is less than 1% and the latter is actually lower, as shown in (Figure 8). This is an encouraging result that demonstrates the feasibility of using DL dose prediction together with NTCP models as an automatic and fast decision support tool for the physicians.

These results are in line with the recent findings by Kouwenberg et al. (Kouwenberg et al. 2021), who achieve an accuracy of 87% for H&N patients eligible for PT by using a Gaussian naïve Bayes classifier. However, unlike our approach, their method does not provide a 3D dose distribution, which can later be used to generate the treatment plan in a fully automated treatment workflow like suggested in (McIntosh et al. 2017). Instead, the Bayes classifier selects a treatment modality beforehand (pre-selection) and the corresponding plan still needs to be generated manually. Another recent study investigates the use of PT and XT dose prediction with DL for treatment selection in pediatric tumors in the abdominal region [9], achieving a similar accuracy (90%). Moreover, the comparison with our work is not straightforward, since their method uses only 20 test patients (versus 55 patients in our case) and the planning of abdominal tumors strongly differs from H&N cases. The study conducted by Hytönen et al. (Hytönen et al. 2022), follows a similar approach of patient referral with the RapidPlan system. However, the authors do not compare the treatment indication obtained with RapidPlan and clinical doses from manual plans, but rather the DVH predicted by the RapidPlan system and the dose generated after optimization. They achieve an accuracy of 81\% but a fair comparison is difficult as we consider the manually generated clinical doses as ground truth

Our method takes only 11s to predict the dose. Transferring a trained model from one center to another is simple. It requires less than 20 megabytes of memory and contains exclusively the values of the optimized DL network parameters. This implies that information about

patients used for training and validation remains inaccessible and fully confidential. Thus, our model represents a way of capturing PT planning expertise and transferring it to centers with fewer cases and less expertise, while preserving patient privacy as there is no need to transfer patient data. Additionally, the required hardware is minimal: a computer to perform the prediction and a way to interface it with the treatment planning system or PACS. This interface should enable a streamline workflow to export the input images (DICOM CT and RTSTRUCT), run the DL model, and import back the predicted doses in DICOM format for visualization and evaluation (i.e. dose metrics, NTCP values). However, due to the addition of the prior dose distribution, the users are required to have a license of a PT plan generation treatment planning system. This requirement reduces the portability of AI-PROTIPP but the gain obtained in the predictions is statistically significant (Appendix 1). Moreover, as this step is entirely automatic, the user should not be trained to generate a PT plan, an expertise rare and expensive.

Centers with a PT machine and a PT planning system could deliver the predicted PT dose with an additional post-processing step: the generation of the corresponding treatment plan (Babier et al. 2020). This process is known as dose mimicking, and consists in using different optimization objectives to find the machine parameters that reproduce the predicted dose. If XT is selected instead, dose mimicking strategies can equally be applied to generate the XT plan. As soon as all functional blocks are pipelined, namely, export/import of images, call of the DL model, treatment choice based on NTCP, and dose mimicking, a final treatment plan can be generated in a few minutes.

Multiple dose mimicking techniques exist, such as the isodose based optimized plan (IBO) (Borderías Villarroel et al. 2020). With IBO, isodose contours are generated from the predicted dose and associated with patient-specific weighted objectives to mimic the prediction. Comparing the planned doses and these mimicked doses in terms of NTCP would shed more light on the applications of our method. In addition, for the generated PT plans, robustness evaluation (Souris et al. 2019) should be performed in order to explore if the robustness of the planned doses is maintained in the mimicked dose predictions. Future research should then be focused on investigating different dose mimicking strategies to ensure a robust and accurate final treatment plan, as well as reducing the dose prediction errors for the small structures (e.g., inferior and middle pharyngeal constrictor muscles). Another perspective of this work is to develop strategies to detect the cases where the prediction has failed. It could raise an alarm and call the dosimetrist to generate a manual plan instead. A way to do that is to estimate the uncertainty associated with the predicted dose, and thus trigger the alarm when the uncertainty becomes too high. This is an ongoing research topic in our center (Vanginderdeuren et al.).

5 Conclusion

Combining DL dose prediction and NTCP models shows promising perspectives to develop an automatic and robust decision support tool for physicians. Our results suggest that DL models can safely be used to predict moderate to severe side effects due to PT or XT treatments (87.4% weighted accuracy) and prevent them by choosing the most appropriate treatment modality. The average cumulative NTCP for each side-effect is not affected by using DL and the time reduction is substantial: around 11 seconds to predict the dose for one

modality rather than hours of semi-manual optimization. Perspectives are to develop strategies to detect cases where the prediction could fail.

Ahn, Sang Hee, Eunsook Kim, Chankyu Kim, Wonjoong Cheon, Myeongsoo Kim, Se Byeong Lee, Young Kyung Lim, et al. 2021. “Deep Learning Method for Prediction of Patient-Specific Dose Distribution in Breast Cancer.” *Radiation Oncology* 16 (1): 154.

Babier, Aaron, Rafid Mahmood, Andrea L. McNiven, Adam Diamant, and Timothy C. Y. Chan. 2020. “The Importance of Evaluating the Complete Automated Knowledge-Based Planning Pipeline.” *Physica Medica: PM: An International Journal Devoted to the Applications of Physics to Medicine and Biology: Official Journal of the Italian Association of Biomedical Physics* 72 (April): 73–79.

Barragán-Montero, A. M., D. Nguyen, and W. Lu. 2019. “Three-dimensional Dose Prediction for Lung IMRT Patients with Deep Neural Networks: Robust Learning from Heterogeneous Beam Configurations.” *Medical*. https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13597.

Barragán-Montero, Ana María, Dan Nguyen, Weiguo Lu, Mu-Han Lin, Roya Norouzi-Kandalan, Xavier Geets, Edmond Sterpin, and Steve Jiang. 2019. “Three-Dimensional Dose Prediction for Lung IMRT Patients with Deep Neural Networks: Robust Learning from Heterogeneous Beam Configurations.” *Medical Physics* 46 (8): 3679–91.

Barragán-Montero, Ana M., Melissa Thomas, Gilles Defraene, Steven Michiels, Karin Haustermans, John A. Lee, and Edmond Sterpin. 2021. “Deep Learning Dose Prediction for IMRT of Esophageal Cancer: The Effect of Data Quality and Quantity on Model Performance.” *Physica Medica: PM: An International Journal Devoted to the Applications of Physics to Medicine and Biology: Official Journal of the Italian Association of Biomedical Physics* 83 (March): 52–63.

Beetz, Ivo, Cornelis Schilstra, Peter van Luijk, Miranda E. M. C. Christianen, Patricia Doornaert, Henk P. Bijl, Olga Chouvalova, Edwin R. van den Heuvel, Roel J. H. M. Steenbakkers, and Johannes A. Langendijk. 2012. “External Validation of Three Dimensional Conformal Radiotherapy Based NTCP Models for Patient-Rated Xerostomia and Sticky Saliva among Patients Treated with Intensity Modulated Radiotherapy.” *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 105 (1): 94–100.

Borderías Villarroel, Elena, Xavier Geets, and Edmond Sterpin. 2020. “Online Adaptive Dose Restoration in Intensity Modulated Proton Therapy of Lung Cancer to Account for Inter-Fractional Density Changes.” *Physics and Imaging in Radiation Oncology* 15 (July): 30–37.

Brodin, N. Patrik, Rafi Kabarriti, Clyde B. Schechter, Mark Pankuch, Vinai Gondi, Shalom Kalnicki, Madhur K. Garg, and Wolfgang A. Tomé. 2021. “Individualized Quality of Life Benefit and Cost-Effectiveness Estimates of Proton Therapy for Patients with Oropharyngeal Cancer.” *Radiation Oncology* 16 (1): 19.

Christianen, Miranda E. M. C., Cornelis Schilstra, Ivo Beetz, Christina T. Muijs, Olga Chouvalova, Fred R. Burlage, Patricia Doornaert, et al. 2012. “Predictive Modelling for Swallowing Dysfunction after Primary (chemo)radiation: Results of a Prospective Observational Study.” *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 105 (1): 107–14.

Fredriksson, Albin, Anders Forsgren, and Björn Hårdemark. 2011. “Minimax Optimization for Handling Range and Setup Uncertainties in Proton Therapy.” *Medical Physics* 38 (3): 1672–84.

Guerreiro, F., E. Seravalli, G. O. Janssens, J. H. Maduro, A. C. Knopf, J. A. Langendijk, B. W. Raaymakers, and C. Kontaxis. 2021. “Deep Learning Prediction of Proton and Photon Dose Distributions for Paediatric Abdominal Tumours.” *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 156 (March): 36–42.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017.

"Densely Connected Convolutional Networks." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–69.
Hytönen, Roni, Marije R. Vergeer, Reynald Vanderstraeten, Timo K. Koponen, Christel Smith, and Wilko F. A. R. Verbakel. 2022. "Fast, Automated, Knowledge-Based Treatment Planning for Selecting Patients for Proton Therapy Based on Normal Tissue Complication Probabilities." *Advances in Radiation Oncology* 7 (4): 100903.
Kandalan, Roya Norouzi, Dan Nguyen, Nima Hassan Rezaeian, Ana M. Barragán-Montero, Sebastiaan Breedveld, Kamesh Namuduri, Steve Jiang, and Mu-Han Lin. 2020. "Dose Prediction with Deep Learning for Prostate Cancer Radiation Therapy: Model Adaptation to Different Treatment Planning Practices." *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 153 (December): 228–35.
Kouwenberg, Jasper, Joan Penninkhof, Steven Habraken, Jaap Zindler, Mischa Hoogeman, and Ben Heijmen. 2021. "Model Based Patient Pre-Selection for Intensity-Modulated Proton Therapy (IMPT) Using Automated Treatment Planning and Machine Learning." *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 158 (May): 224–29.
"Landelijk Indicatieprotocol Protonentherapie Hoofdhals Versie 2.2, 2019." 2019. August 15, 2019. https://nvro.nl/images/documenten/rapporten/2019-08-15__Landelijk_Indicatieprotocol_Protonentherapie_Hoofdhals_v2.2.pdf.
Langendijk, Johannes A., Philippe Lambin, Dirk De Ruysscher, Joachim Widder, Mike Bos, and Marcel Verheij. 2013. "Selection of Patients for Radiotherapy with Protons Aiming at Reduction of Side Effects: The Model-Based Approach." *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 107 (3): 267–73.
Leroy, Roos, Cindy De Gendt, Sabine Stordeur, Viki Schillemans, Leen Verleye, Geert Silversmit, Elizabeth Van Eycken, et al. 2019. "Head and Neck Cancer in Belgium: Quality of Diagnostic Management and Variability Across Belgian Hospitals Between 2009 and 2014." *Frontiers in Oncology* 9 (October): 1006.
Liu, Runxin, Jingfeng Bai, Jingjie Zhou, Kang Zhang, and Cheng Ni. 2020. "A Feasibility Study for Predicting 3D Radiotherapy Dose Distribution of Lung VMAT Patients." In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 1304–8.
McIntosh, Chris, Mattea Welch, Andrea McNiven, David A. Jaffray, and Thomas G. Purdie. 2017. "Fully Automated Treatment Planning for Head and Neck Radiotherapy Using a Voxel-Based Dose Prediction and Dose Mimicking Method." *Physics in Medicine and Biology* 62 (15): 5926–44.
Nelms, Benjamin E., Greg Robinson, Jay Markham, Kyle Velasco, Steve Boyd, Sharath Narayan, James Wheeler, and Mark L. Sobczak. 2012. "Variation in External Beam Treatment Plan Quality: An Inter-Institutional Study of Planners and Planning Systems." *Practical Radiation Oncology* 2 (4): 296–305.
Nguyen, Dan, Xun Jia, David Sher, Mu-Han Lin, Zohaib Iqbal, Hui Liu, and Steve Jiang. 2019. "3D Radiotherapy Dose Prediction on Head and Neck Cancer Patients with a Hierarchically Densely Connected U-Net Deep Learning Architecture." *Physics in Medicine and Biology* 64 (6): 065020.
Nguyen, D., X. Jia, D. Sher, M. H. Lin, and Z. Iqbal. 2019. "3D Radiotherapy Dose Prediction on Head and Neck Cancer Patients with a Hierarchically Densely Connected U-Net Deep Learning Architecture." *Physics in Medicine and Biology*. https://iopscience.iop.org/article/10.1088/1361-6560/ab039b/meta.
Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–41. Springer International Publishing.
Song, Ying, Junjie Hu, Yang Liu, Haiyun Hu, Yang Huang, Sen Bai, and Zhang Yi. 2020. "Dose Prediction Using a Deep Neural Network for Accelerated Planning of

Rectal Cancer Radiotherapy.” *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 149 (August): 111–16.
Souris, Kevin, Ana Barragan Montero, Guillaume Janssens, Dario Di Perri, Edmond Sterpin, and John A. Lee. 2019. “Monte Carlo Methods to Comprehensively Evaluate the Robustness of 4D Treatments in Proton Therapy.” *Medical Physics* 46 (10): 4676–84.
Tambas, Makbule, Hans Paul van der Laan, Arjen van der Schaaf, Roel J. H. M. Steenbakkers, and Johannes Albertus Langendijk. 2022. “A Decision Support Tool to Optimize Selection of Head and Neck Cancer Patients for Proton Therapy.” *Cancers* 14 (3). https://doi.org/10.3390/cancers14030681.
Tambas, Makbule, Roel J. H. M. Steenbakkers, Hans P. van der Laan, Atje M. Wolters, Roel G. J. Kierkels, Dan Scandurra, Erik W. Korevaar, et al. 2020. “First Experience with Model-Based Selection of Head and Neck Cancer Patients for Proton Therapy.” *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 151 (October): 206–13.
Vanginderdeuren, and Huet-Dastarac. “Estimating Uncertainty in Radiation Oncology Dose Prediction with Dropout and Bootstrap in U-Net Models.” *On Artificial Neural ….* https://dial.uclouvain.be/pr/boreal/object/boreal%3A257291/datastream/PDF_01/view.

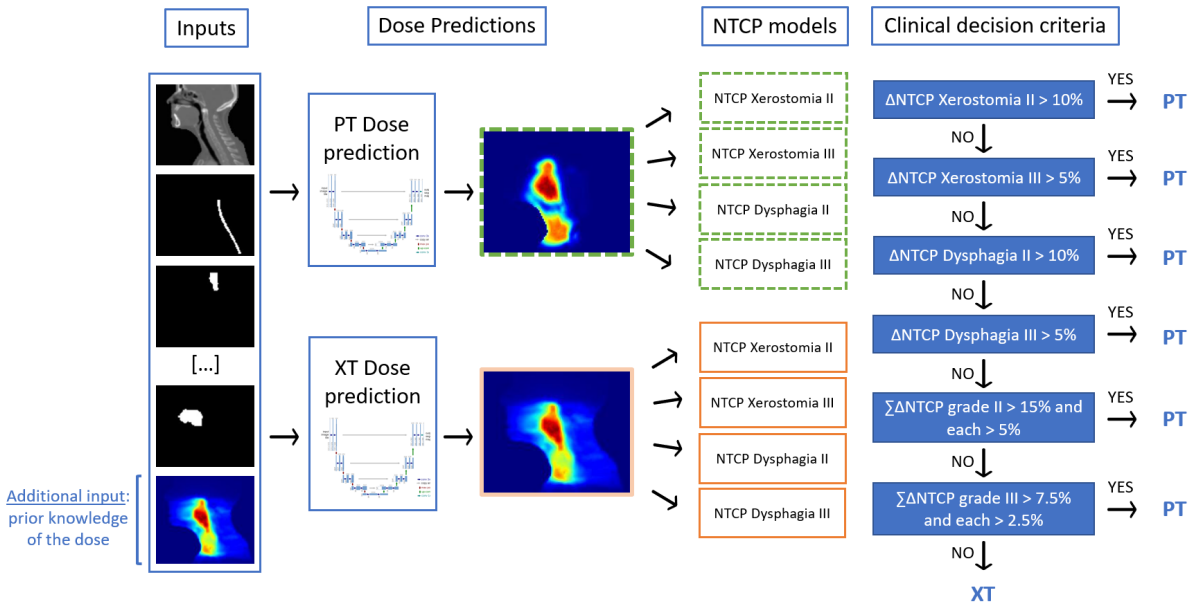


Figure 1: AI-PROTIPP workflow from the input to the AI models to the treatment selection.

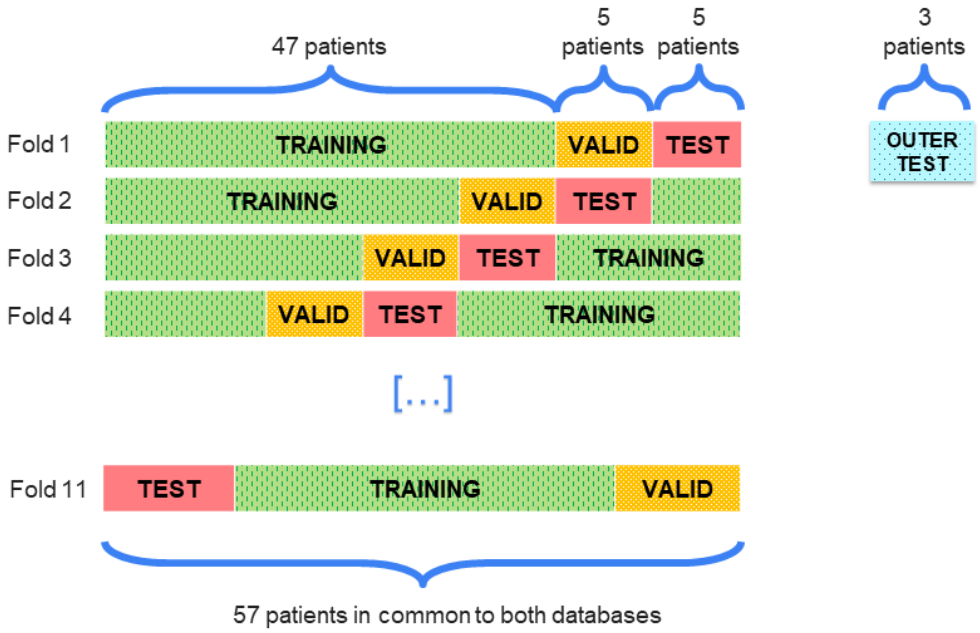


Figure 2: 11-fold nested cross-validation method to obtain the predicted doses for 55 patients of the database. 3 patients were set aside in the outer test set to ensure the validity of the method. For each fold, we select 5 patients for the test, 5 patients for the validation and the remaining 47 patients for the training set. The predictions for the test set were stored before retraining a model with the subsequent partition of patients.

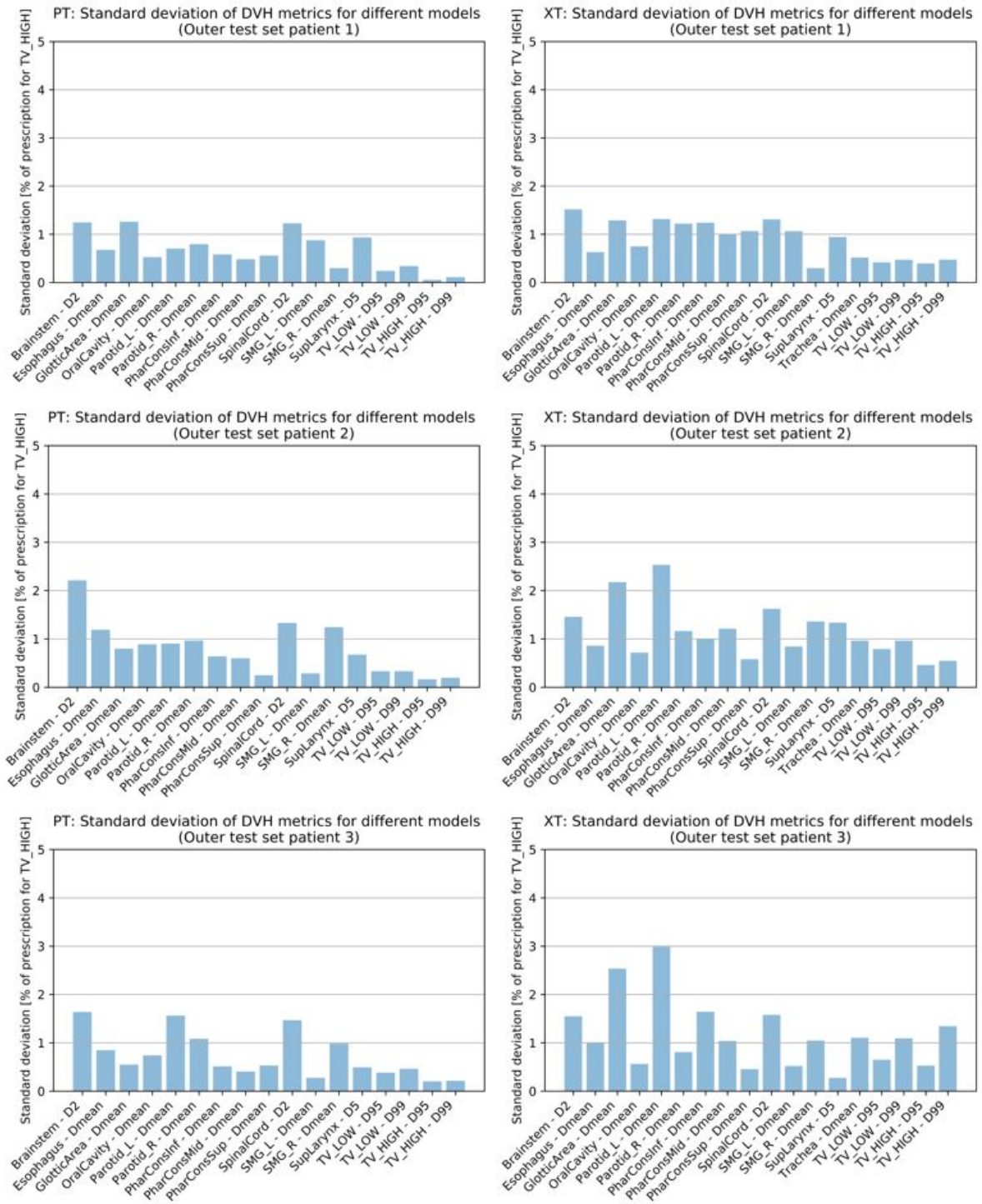


Figure 3: Standard deviation of the DVH metrics computed from the prediction of the 11 models on the 3 patients from the outer test set. Only the remaining 57 patients were used for training, validation and test. Glossary of structure labels on the horizontal axis can be found in appendix A2.

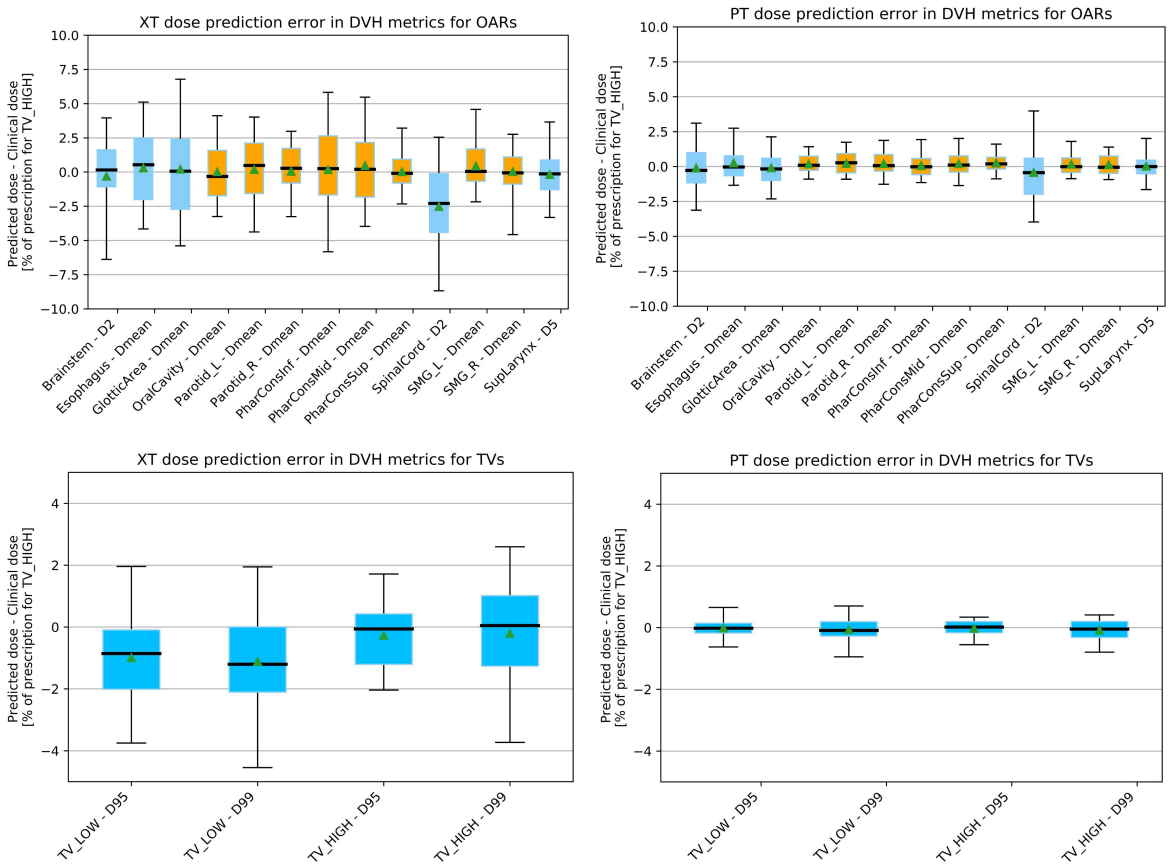


Figure 4: Boxplots displaying performance of the DL dose prediction models over the 55 patients with, on the left, the network predicting the dose for XT treatment and, on the right, the network predicting the dose for the PT treatment. Bottom subfigures display the metrics D95 and D99 for the tumor volumes, top ones display the most clinically relevant DVH metric among Dmean, D5 and D2 for the OARs. In orange are highlighted the organs which Dmean are inputted to the NTCP models. The boxes, horizontal lines and whiskers show the quartiles, medians, and the [5, 95] percentile range of the prediction error, respectively. The green triangles represent the mean. Glossary of structure labels on the horizontal axis can be found in appendix A2.

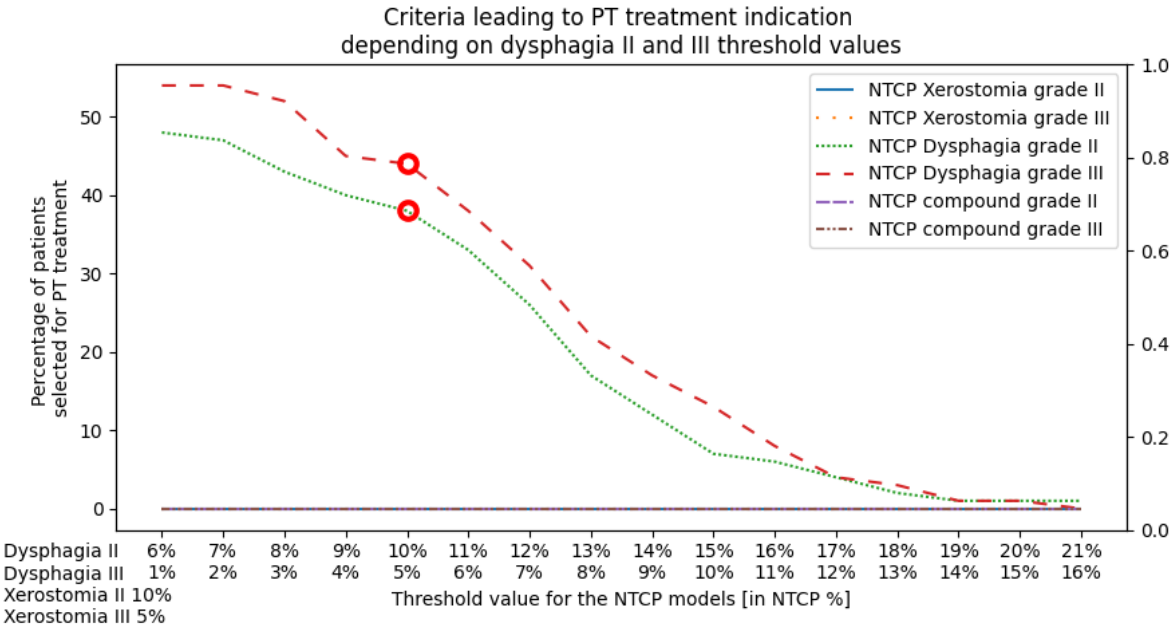


Figure 5: Criteria leading to a PT indication along different NTCP difference thresholds. The LIPPv2.2 thresholds as fixed by the Health Council of the Netherlands are indicated with the red circles.

Figure 6: Display of the NTCP differences for predicted and clinical doses for the falsely selected patients: either false positive or false negative.

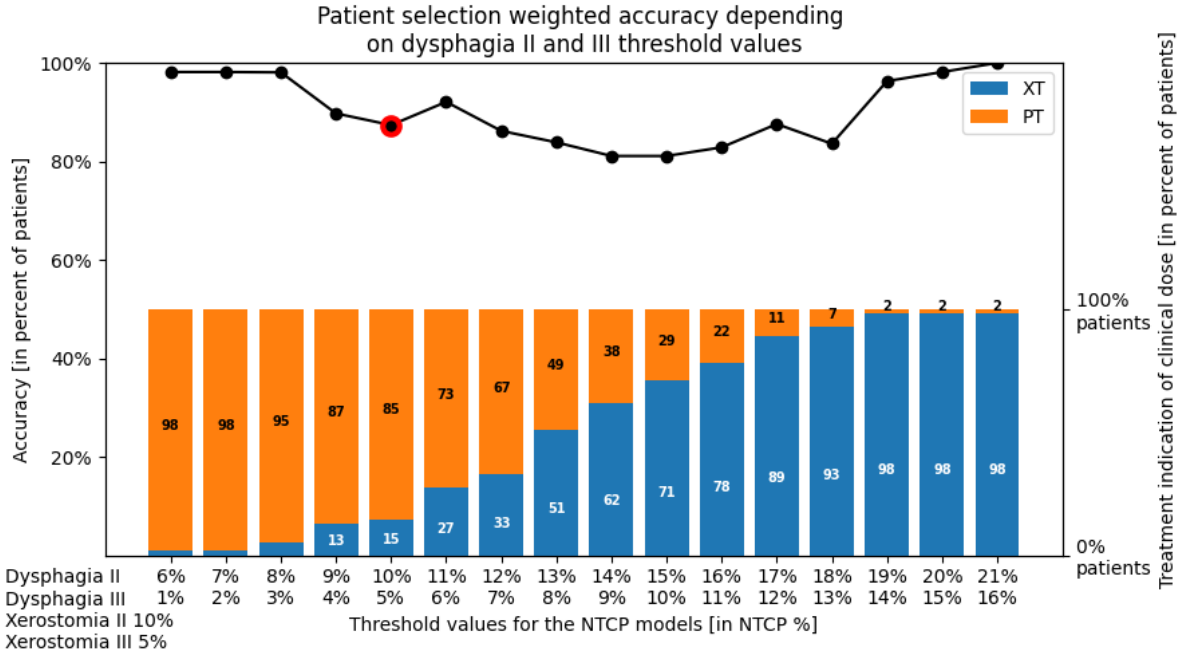


Figure 7: Variation of the thresholds selecting patients for PT. The bar chart represents the partition patients selected for PT in orange and XT in blue with the clinically planned doses. The line chart displays the weighted accuracy and the set of default thresholds set by the Health Council of the Netherlands is outlined in red.

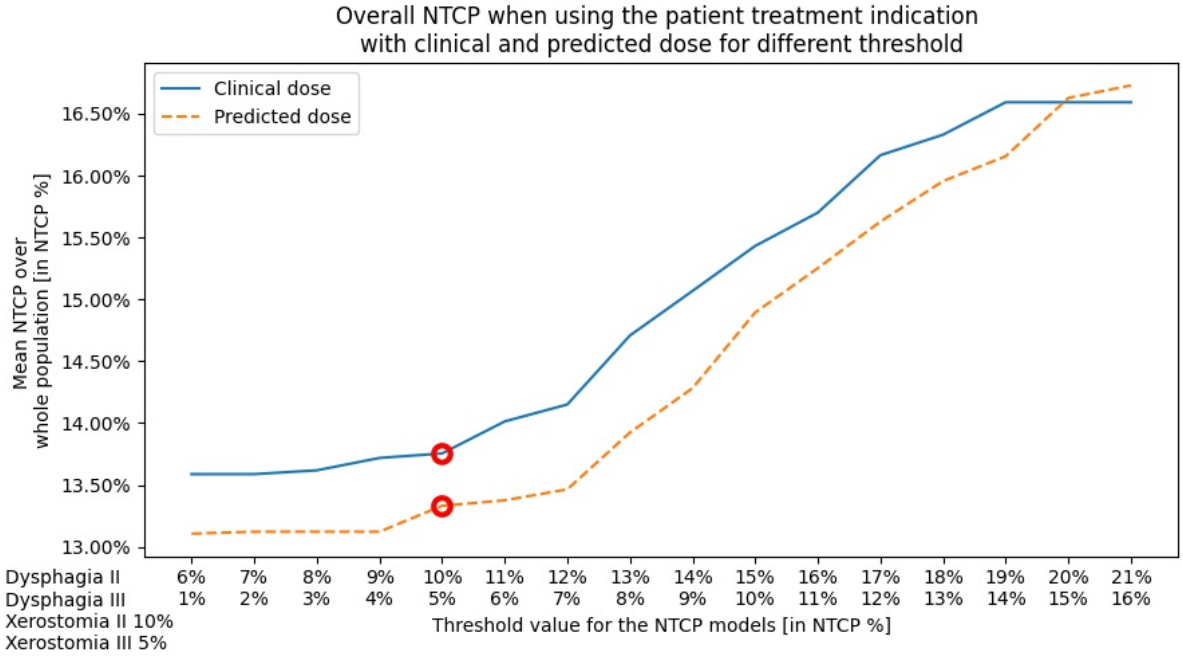
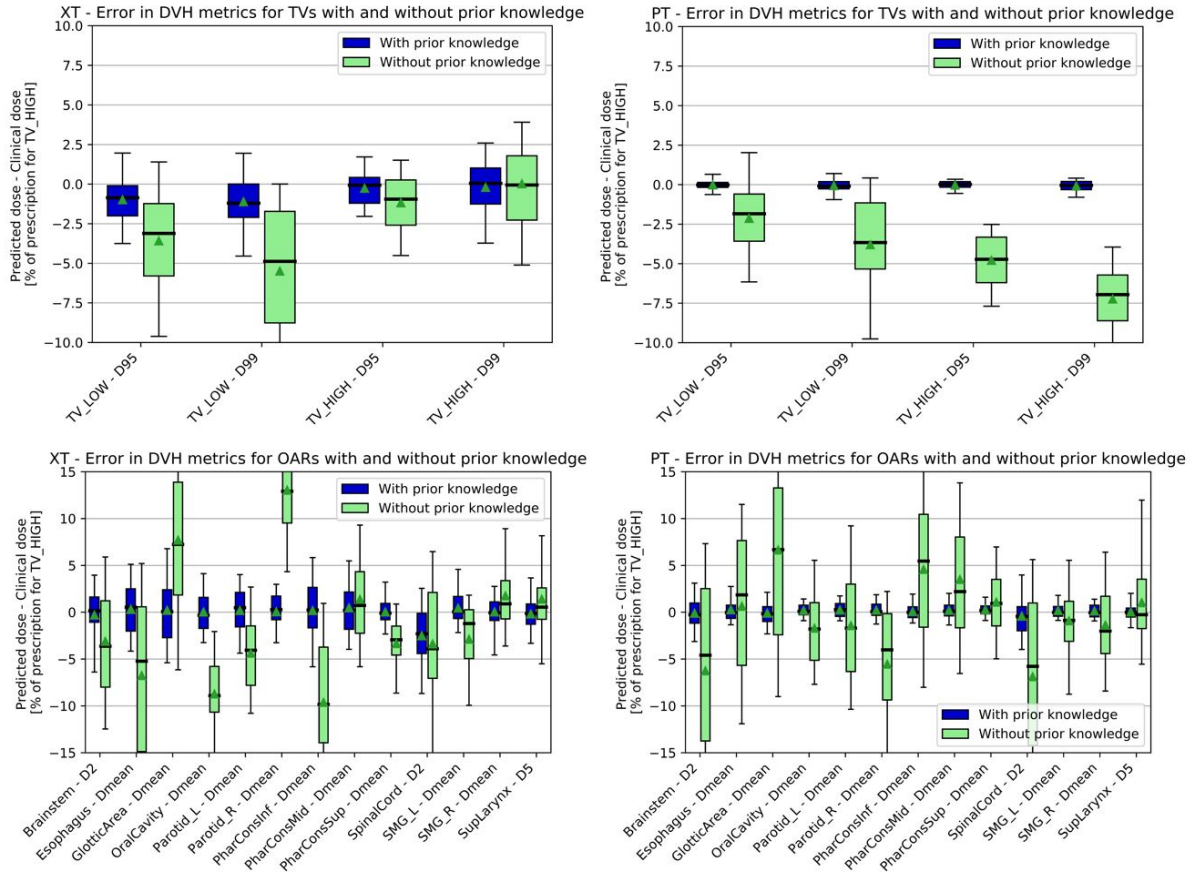


Figure 8: Mean NTCP over the clinical plans of the 55 patients following the treatment indicated for different threshold values. The set of default thresholds set by the Health Council of the Netherlands is outlined in red.

Dmean of OAR inputted in the NTCP models	Wilcoxon Rank Sum Test p-values	
	PT	XT
Oral cavity	0.104	0.512
Left parotid	0.287	0.384
Right parotid	0.828	0.513
Inferior pharyngeal constrictor muscle	0.215	0.357
Middle pharyngeal constrictor muscle	0.059	0.947
Superior pharyngeal constrictor muscle	0.234	0.854
Left submandibular gland	0.555	0.727
Right submandibular gland	0.451	0.497

Table 1: P-values paired-samples Wilcoxon Test between predicted and clinical doses. Statistical significance is achieved if the p-value is inferior to 0.05.



Appendix A1: Comparison of the error between the clinical and the predicted doses when the prior knowledge is inputted or not in the network.

Name of structures in figures	**Description**
Brainstem	Brainstem
Esophagus	Esophagus
GlotticArea	Glottic Area
Parotid_L	Left Parotid
Parotid_R	Right Parotid
PharConstrInf	Inferior pharyngeal constrictor muscle
PharConstrMid	Middle pharyngeal constrictor muscle
PharConstrSup	Superior pharyngeal constrictor muscle
SpinalCord	Spinal Cord
SMG_L	Left submandibular gland
SMG_R	Right submandibular gland

SupLarynx	Supraglottic larynx
TV_LOW	Target volume with prescription: 54.25Gy
TV_HIGH	Target volume with prescription: 70.00Gy

Appendix A2: Glossary of structures.