

**PTV-based VMAT vs. robust IMPT for head-and-neck cancer**

**A probabilistic uncertainty analysis of clinical plan evaluation with the Dutch model-based selection**

Rojo-Santiago, Jesús; Korevaar, Erik; Perkó, Zoltán; Both, Stefan; Habraken, Steven J.M.; Hoogeman, Mischa S.

**DOI**

[10.1016/j.radonc.2023.109729](https://doi.org/10.1016/j.radonc.2023.109729)

**Publication date**

2023

**Document Version**

Final published version

**Published in**

Radiotherapy and Oncology

**Citation (APA)**

Rojo-Santiago, J., Korevaar, E., Perkó, Z., Both, S., Habraken, S. J. M., & Hoogeman, M. S. (2023). PTV-based VMAT vs. robust IMPT for head-and-neck cancer: A probabilistic uncertainty analysis of clinical plan evaluation with the Dutch model-based selection. *Radiotherapy and Oncology*, 186, Article 109729. <https://doi.org/10.1016/j.radonc.2023.109729>

**Important note**

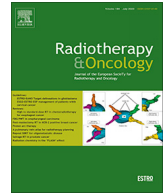
To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



## Original Article

# PTV-based VMAT vs. robust IMPT for head-and-neck cancer: A probabilistic uncertainty analysis of clinical plan evaluation with the Dutch model-based selection



Jesús Rojo-Santiago<sup>a,b,1,2,\*</sup>, Erik Korevaar<sup>c</sup>, Zoltán Perkó<sup>d</sup>, Stefan Both<sup>c</sup>, Steven J.M. Habraken<sup>a,b</sup>, Mischa S. Hoogeman<sup>a,b</sup>

<sup>a</sup>Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, the Netherlands; <sup>b</sup>Department of Medical Physics & Informatics, HollandPTC, Delft; <sup>c</sup>Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen; and <sup>d</sup>Delft University of Technology, Department of Radiation Science and Technology, Delft, the Netherlands

## ARTICLE INFO

## Article history:

Received 12 January 2023  
Received in revised form 9 May 2023  
Accepted 23 May 2023  
Available online 8 June 2023

## Keywords:

Intensity Modulated Proton Therapy  
Volumetric Modulated Arc Therapy  
Polynomial Chaos Expansion  
Probabilistic planning  
Head-and-neck Cancer

## ABSTRACT

**Background and purpose:** In the Netherlands, head-and-neck cancer (HNC) patients are referred for proton therapy (PT) through model-based selection (MBS). However, treatment errors may compromise adequate CTV dose. Our aims are: (i) to derive probabilistic plan evaluation metrics on the CTV consistent with clinical metrics; (ii) to evaluate plan consistency between photon (VMAT) and proton (IMPT) planning in terms of CTV dose iso-effectiveness and (iii) to assess the robustness of the OAR doses and of the risk toxicities involved in the MBS.

**Materials and methods:** Sixty HNC plans (30 IMPT/30 VMAT) were included. A robustness evaluation with 100,000 treatment scenarios per plan was performed using Polynomial Chaos Expansion (PCE). PCE was applied to determine scenario distributions of clinically relevant dosimetric parameters, which were compared between the 2 modalities. Finally, PCE-based probabilistic dose parameters were derived and compared to clinical PTV-based photon and voxel-wise proton evaluation metrics.

**Results:** Probabilistic dose to near-minimum volume  $v = 99.8\%$  for the CTV correlated best with clinical PTV- $D_{98\%}$  and  $VW_{min-D_{98\%,CTV}}$  doses for VMAT and IMPT respectively. IMPT showed slightly higher nominal CTV doses, with an average increase of 0.8 GyRBE in the median of the  $D_{99.8\%,CTV}$  distribution. Most patients qualified for IMPT through the dysphagia grade II model, for which an average NTCP gain of 10.5 percentage points (%-point) was found. For all complications, uncertainties resulted in moderate NTCP spreads lower than 3 p.p. on average for both modalities.

**Conclusion:** Despite the differences between photon and proton planning, the comparison between PTV-based VMAT and robust IMPT is consistent. Treatment errors had a moderate impact on NTCPs, showing that the nominal plans are a good estimator to qualify patients for PT.

© 2023 The Author(s). Published by Elsevier B.V. Radiotherapy and Oncology 186 (2023) 109729 This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Radiation therapy (RT) in combination with chemotherapy is widely used in the treatment of head-and-neck cancer (HNC) patients. Highly conformal RT techniques such as intensity-modulated radiation therapy (IMRT) and volumetric-modulated arc therapy (VMAT) have been used to precisely deliver the dose [1]. More recently, intensity-modulated proton therapy (IMPT) has also been introduced clinically. Compared to RT, its physics properties allow achieving similar dose conformality around the target with an additional improvement on healthy tissue sparing

[2]. However, IMPT also comes with higher costs and limited capacity.

In the Netherlands, a standardized model-based selection (MBS) is used to select HNC patients for proton therapy (PT) [3–5]. The Dutch MBS for HNC patients consists of a plan comparison between an IMPT and a VMAT plan, in which a patient is selected for IMPT if sufficient dosimetric benefit is achieved in modelled normal tissue complication probabilities (NTCPs). The plan comparison is based on the assumption of plan iso-effectiveness in the clinical target volume (CTV), which means an equivalence in CTV dose and tumor control probability (TCP) has to be achieved for both VMAT and IMPT. Treatment errors and their impact on delivered dose are, however, different for photons and protons, and can compromise this iso-effectiveness. First, IMPT is more

\* Corresponding author at: Doctor Molewaterplein 40, 3015GD Rotterdam, the Netherlands.

E-mail address: [j.rojosantiago@erasmusmc.nl](mailto:j.rojosantiago@erasmusmc.nl) (J. Rojo-Santiago).

<sup>1</sup> Postal address: P.O. box 2040.

<sup>2</sup> Street address: Doctor Molewaterplein 40, 3015GD, Rotterdam, The Netherlands.

sensitive to beam- and patient-alignment (geometrical) errors and to anatomical variations than VMAT [6]. Second, IMPT is also subject to uncertainties in the stopping-power prediction (SPP), i.e., the range [7]. Moreover, geometrical and range errors also have an impact on the dose delivered to the organs-at-risk (OARs), and, subsequently, can impact the estimated NTCP values.

To mitigate uncertainties, conventional RT and PT uses fundamentally different approaches. For VMAT, a planning target volume (PTV) is used, assuming dose invariance against shifts (static dose cloud approximation) [8]. For IMPT, a scenario-based robust optimization and a voxel-wise evaluation are used instead [9–11]. Both strategies enlarge the treated volume around the CTV to ensure adequate target coverage, resulting in increased doses to normal tissues. Clinical plan robustness evaluations are usually based on these enlarged volumes i.e., PTV-based and voxel-wise metrics for VMAT and IMPT, respectively, while uniform, more comprehensive and probabilistic plan evaluations on the CTV are lacking. Due to these differences, it is unknown if plan iso-effectiveness on the CTV between the 2 modalities is actually achieved in clinical practice.

In this study, we accurately determined whether iso-effectiveness for the CTV was achieved between the 2 treatment modalities within the Dutch MBS. We used polynomial chaos expansion (PCE) to fast and accurately model the dependence of delivered dose on the geometrical and range errors. With PCE, we simulated 100.000 complete fractionated treatments per plan, obtaining probabilistic distributions of relevant dosimetric parameters for the primary and elective CTVs and main OARs [12–14]. First, we derived and analyzed consistent probabilistic CTV dose parameters to give a more realistic, physical and statistical meaning to PTV-based and voxel-wise dose clinical evaluation metrics. Second, we evaluated consistency and robustness between VMAT and IMPT in terms of CTV dose, and subsequently, in terms of TCPs. Finally, we assessed the sensitivity of OAR doses and NTCP values against geometrical and range errors to investigate their impact on the MBS for PT patient selection.

## Methods and materials

### Patient data

Thirty oropharyngeal HNC patients, treated at the UMCG with VMAT (10/30) or IMPT (20/30), were included. For each patient, a VMAT and an IMPT plan were available from the planning comparison, leading to a total of 60 plans. Prescribed doses ( $D_{\text{pres}}$ ) were 70 Gy for the primary CTV ( $\text{CTV}_{70,00}$ ) and 54.25 Gy to the elective lymph nodes ( $\text{CTV}_{54,25}$ ), both delivered in 35 fractions. For VMAT, dose was prescribed to the PTV:  $\text{PTV-D}_{98\%} \geq 95\% D_{\text{pres}}$ , while, for IMPT, dose was prescribed to the voxel-wise near-minimum (VWmin) CTV dose:  $\text{VWmin-D}_{98\%,\text{CTV}} \geq 94\% D_{\text{pres}}$  [11]. To avoid hot spots in the target, an additional constraint was included for the PTV near-maximum dose:  $\text{PTV-D}_{2\%} \leq 107\% D_{\text{pres}}$  for VMAT and on the voxel-wise near-maximum (VWmax) CTV dose:  $\text{VWmax-D}_{1\text{cc},\text{CTV}} \leq 78 \text{ GyRBE}$  for IMPT. A constant radiobiological effectiveness (RBE) of 1.1 was assumed for IMPT relative to VMAT.

To facilitate gross tumour volume (GTV) and OARs delineation, a T2-weighted sequence of magnetic resonance imaging (MRI) and a positron emission tomography (PET) scan were acquired per patient. They were rigidly registered with a single-energy planning computed tomography (SECT) scan, with  $0.98 \times 0.98 \times 2 \text{ mm}^3$  CT resolution. The median GTV was 13.49 cc (range 1.37–117.05 cc). Main relevant OARs i.e., the parotid and submandibular glands, spinal cord, extended oral cavity and pharyngeal constrictor muscles (PCM) were delineated according to the international consensus guidelines for CT-based delineation of OARs in the head-and-neck region [15].

### Treatment planning

Both VMAT and IMPT treatment plans were available in RayStation (version 10B, RaySearch, Sweden). For VMAT, dual 6MV arcs were used. Dose was calculated using a collapsed cone (CC) dose algorithm. To handle geometrical errors, a PTV-based optimization and evaluation was performed using an isotropic CTV-PTV margin of 3 mm. When the PTV was adequately covered, the dose optimization for the OARs was guided by NTCP-based cost functions [16].

For IMPT, most plans were generated using an initial arrangement of 4 beam angles, with 2 anterior ( $40^\circ$  and  $320^\circ$ ) and 2 posterior ( $160^\circ$  and  $200^\circ$ ) coplanar oblique beams. In 3 patients, additional beams were used to increase robustness against anatomical changes. Subsequently, beam angles were manually optimized per patient to better spare the PCM and the parotid glands. Posterior beams were split into range shifter and non-range shifter fields, resulting in 5 to 7 beams per plan. Dose of the clinical IMPT plans was calculated using a Monte Carlo (MC) dose engine with 1% MC uncertainty, but also were subsequently re-evaluated with a MC noise of 0.1%. Robust minimax optimization [9,10] was used to handle geometrical and range errors. Based on the clinical evaluation of the first cohort of patients [17], a setup robustness (SR) setting of 3 mm was used. For the relative SPP error, a range robustness (RR) setting of 3% was assumed based on literature [18] and validated in-house [19,20]. Finally, a robustness evaluation was performed in line with [11]. CTV coverage and OARs doses were respectively ensured using VWmin and VWmax dose distributions from 28 error scenarios. They consisted of the combination of 14 geometrical shifts and 2 range errors based on the SR (3 mm) and RR (3%) clinically used.

To reduce inter-patient and inter-modality variations, all treatment plans were normalized to their clinical  $\text{VWmin-D}_{98\%,\text{CTV}}$  ( $G_s$ , IMPT) and  $\text{PTV-D}_{98\%}$  ( $G_s$ , VMAT) values respectively. For IMPT, two normalization levels were considered: (1) 94%, applied in Sections 2.5 and 2.6 to assess treatment plan differences according to clinical protocol and (2) 94.5%, applied in Section 2.4 according to our adjusted protocol based on a re-calibration, in order to consistently derive probabilistic dose metrics on the CTV. The normalization was based on the geometrical scaling factor  $G_s$  that is described in the [Supplementary Material \(SM\)](#) [SM, Section S1].

### Polynomial chaos expansion (PCE): PCE-based robustness evaluation

The impact of geometrical and range errors on CTV and OARs doses was modelled with Polynomial Chaos Expansion (PCE). PCE provides a computationally efficient patient- and treatment plan-specific analytical model of the dose engine. PCE approximates the dose  $D_i$  of each voxel  $i$  affected by a geometrical shift  $\vec{\xi} = (\xi_x, \xi_y, \xi_z)$  and a relative range error  $\rho$  as the series expansion  $D_i(\vec{\xi}, \rho) = \sum_{k=0}^p a_{i,k} \Psi_k(\vec{\xi}, \rho)$ , with expansion coefficients  $\{a_{i,k}\}$  and multi-dimensional Hermite polynomials  $\Psi_k(\vec{\xi}, \rho)$  [12–14]. It enables the evaluation (PCE-based robustness evaluation) of 100,000 complete fractionated treatments with proper statistical weighting. A complete VMAT fractionated treatment is simulated by drawing (i) a fixed systematic geometrical error ( $\Sigma$ ) for the complete treatment and (ii) a different random error ( $\sigma$ ) for each fraction, both characterized by the standard deviations (1 SD) of Gaussian error distributions. For the IMPT treatments, an additional (iii) systematic range ( $\rho$ ) error (1 SD) for each treatment is included.

We considered two cases of  $\Sigma$  and  $\sigma$  errors. For the derivation of the probabilistic plan evaluation metrics, values of  $\Sigma = 0.92 \text{ mm}$  and  $\sigma = 1.00 \text{ mm}$  for the systematic and random geometrical errors (1 SD) were assumed, consistent with a  $M = 2.5\Sigma + 0.7\sigma = 3 \text{ mm}$

margin based on van Herk's recipe and clinical experience (error set I). To assess treatment plan differences between VMAT and IMPT, a second combination of systematic and random geometrical errors were determined from clinical measurements of beam- and patient-alignment errors and machine specifications, as tabulated in Table 1 (error set II). From the patient residual setup analysis, systematic values were acquired as the standard deviation (1 SD) of the residual mean setup errors, while the random components were obtained as the root-mean-square value of the residual setup standard deviations. Range uncertainties were based on literature [21].

Patient- and plan-specific PCE models were constructed for both treatment modalities using the clinical RayStation CC/MC dose engine. To further improve PCE accuracy for the IMPT plans, the PCE models were constructed using a MC noise of 0.1% instead of the clinical 1%. Four PCE-based robustness evaluations were done per patient, two for the VMAT and other two for the IMPT plan, depending on the error set used. Patient and treatment plan-specific probability distributions of relevant dosimetric parameters across the sampled treatments (scenario distributions) were derived for the CTV<sub>70.00</sub>, CTV<sub>54.25</sub> and main relevant OARs. A flowchart of the followed methodology is described in Fig. 1.

#### Probabilistic plan evaluation metrics to assess CTV coverage

In clinical practice, the lack of consistency between plan evaluation metrics hinders a fair comparison of the robustness between photon and proton treatment plans. Additionally, the volume  $v$  of the CTV that is clinically covered by 95%  $D_{pres}$  is not well defined since the clinical goal on the PTV was relaxed from the point minimum dose (PTV- $D_{100\%}$ ), which aimed a CTV coverage of 100% and was used in the derivation of van Herk's margin recipe [8], to the near-minimum dose (PTV- $D_{98\%}$ ) [22].

To this end, we determined plan evaluation metrics ( $D_{v,CTV}$ ) to probabilistically assess target dose adequacy, and we compared them with clinical PTV-based and voxel-wise metrics. First, we compared the PTV- and VWmin-based prescription for VMAT and found the equivalence to PTV-based prescription if VWmin-based is prescribed to 94.5% of reference dose [SM, Section S2], in line with [11]. Second, we scaled the IMPT plans to the 94.5% VWmin-based prescription. Third, PCE-based robustness evaluations were done assuming geometrical errors (1 SD) consistent with a  $M = 3$  mm margin (error set I) based on van Herk's recipe ( $\Sigma = 0.92$  mm,  $\sigma = 1.00$  mm). From the results, we derived the volume  $v$  of the CTV that correlates best with clinical VWmin- $D_{98\%,CTV}$  and PTV- $D_{98\%}$  values, satisfying that, in 90% of the sampled treatments, at least 95%  $D_{pres}$  on average for all patients was achieved in the CTVs. This volume was then used in the evaluation of the CTV dose iso-effectiveness.

#### Evaluation of CTV dose iso-effectiveness

CTV dose iso-effectiveness was evaluated with a comparison of relevant CTV dose parameters and modelled TCP values between the 2 treatment modalities. For this part of the analysis, PCE-based robustness evaluations were done using the clinical geometrical systematic and random errors tabulated in Table 1 (error set II) and IMPT plans scaled to the 94% VWmin-based prescription [SM, Section S1].

Target coverage for both CTVs was compared based on the probabilistic  $D_{v,CTV}$  metrics derived in Section 2.4. Dose conformality of the plans was determined according to a conformity index (CI) value, defined as the ratio between the volume covered by the prescribed isodose (VRI) and the target volume (TV):  $CI = VRI/TV$ . Inhomogeneity of the dose was evaluated by an homo-

geneity index (HI) value, defined as the ratio between  $D_{95\%}/D_{5\%}$  for the CTV<sub>70.00</sub> and CTV<sub>54.25</sub>.

TCP modelling was based on the model by Luhr et. al. [23]. In the TCP model, the target volume is split up in 3 different non-overlapping subvolumes (GTV, CTV<sub>70.00</sub> and CTV<sub>54.25</sub>) with different response to dose [23–25]. TCP scenario distributions for the GTV, disjoint CTV<sub>70.00</sub> and disjoint CTV<sub>54.25</sub> were calculated from the dose-volume histograms of the treatments sampled in each of the regions. The total TCP (TCP<sub>T</sub>) was calculated as:  $TCP_T = TCP_{GTV} \times TCP_{CTV,70} \times TCP_{CTV,54.25}$ . For further details, we refer to [SM, Section S3].

#### Robustness of the MBS for PT patient selection

To assess the robustness of the MBS approach, OARs doses were evaluated with PCE using the clinical geometrical and range errors from Table 1 (error set II) and IMPT plans were scaled to the 94% VWmin-based prescription [SM, Section S1]. To this end, scenario  $D_{mean}$  distributions for the parotids, submandibular glands, the extended oral cavity and the PCM were determined and compared between the 2 modalities.

NTCP models, clinically used for IMPT patient selection [3–5], were also used to facilitate a plan comparison between VMAT and IMPT in terms of tissue toxicities. In this study, we considered the NTCP models for two different toxicities as used for MBS in current clinical practice: (i) patient-rated moderate (grade II) and severe (grade III) xerostomia toxicities and (ii) physician-rated grade II and III dysphagia toxicities. They are based on the  $D_{mean}$  doses from: (i) the parotids and submandibular glands for the xerostomia model and (ii) the external oral cavity and the PCM for the dysphagia model. If the  $\Delta NTCP$  ( $\Delta NTCP = NTCP_{VMAT} - NTCP_{IMPT}$ ) exceeds a threshold for a certain toxicity e.g., 10% and 5% for grade II and grade III complications, respectively, the patient would qualify for IMPT instead of VMAT. NTCP spreads were calculated as the difference between the 5th and the 95th percentile of the scenario NTCP distribution.

#### Statistical analysis

Statistical significance of the dosimetric differences between treatment modalities was assessed using the Wilcoxon signed rank test ( $p < 0.05$ ).

#### Results

Based on the calibration for the probabilistic  $D_{v,CTV}$  metric [SM, Section S4], the 10<sup>th</sup> percentile of  $D_v$  for  $v = 99.8\%$  ( $D_{99.8\%,CTV}$ ) correlated best with clinical VWmin- $D_{98\%,CTV}$  for IMPT and PTV- $D_{98\%}$  for VMAT values. As depicted in Fig. 2, the 10<sup>th</sup> percentile of the scenario  $D_{99.8\%,CTV}$  distributions for both CTVs (CTV<sub>70.00</sub> and CTV<sub>54.25</sub>) resulted in an underestimation of 0.2% compared to the clinical PTV- $D_{98\%}$  for VMAT, while, for IMPT, a slightly higher overestimation of 0.5% of the clinical VWmin- $D_{98\%,CTV}$  was found. For symmetry, a volume  $v = 0.2\%$  ( $D_{0.2\%,CTV}$ ) is suggested to evaluate maximum CTV dose values, which showed weaker correlations with the clinical VWmax- $D_{2\%,CTV}$  ( $R^2 > 0.62$ ) and PTV- $D_{2\%}$  ( $R^2 > 0.94$ ). The correlations for the  $D_{0.2\%,CTV}$  are displayed in the [SM, Section S5].

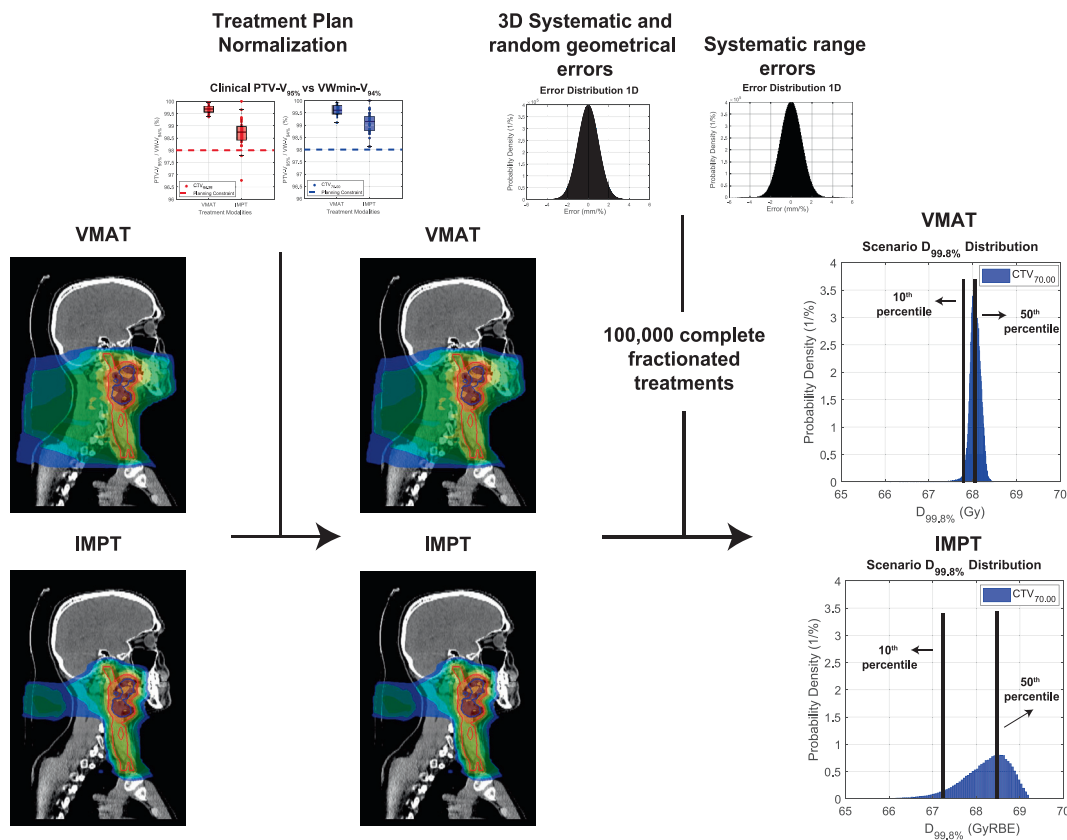
To assess CTV dose iso-effectiveness and robustness of the MBS, median values of relevant dosimetric parameters over all patients are tabulated in Table 2. They are reported with a MC uncertainty of 0.1% to avoid sampling differences between the clinical dose engine and PCE-based evaluations.

Considering the clinical uncertainties from Table 1, a higher CTV near minimum dose was found for IMPT, resulting in a median population  $D_{99.8\%,CTV}$  of 67.6 (10th percentile: 66.8) GyRBE and of

**Table 1**

Geometrical and range uncertainties taken into account for the PCE robustness evaluation. Systematic and random geometrical errors were determined from clinical measurements of beam- and patient-alignment errors and machine specifications, which were split into a systematic and a random component. This set of errors is referred to as error set II throughout the paper.

Range uncertainties (%)						
Uncertainty	Systematic ± 1 SD					
<b>SPP Error</b>	<b>0.0 ± 1.5</b>					
<b>Geometrical uncertainties (mm)</b>						
Uncertainty	Systematic ± 1 SD			Random ± 1 SD		
<b>Isocentric inaccuracies</b>	Left-Right	Dorsoventral	Craniocaudal	Left-Right	Dorsoventral	Craniocaudal
CT Isocenter	± 0.5	± 0.5	± 0.5	± 0.5	± 0.5	± 0.5
Isocenter Gantry	± 0.2	± 0.1	± 0.1	± 0.1	± 0.3	± 0.1
Couch				± 0.5	± 0.5	± 0.5
Online Matching				± 0.5	± 0.5	± 0.5
MR registration	± 0.5	± 0.5	± 0.5			
<b>Patient setup errors</b>	Left-Right	Dorsoventral	Craniocaudal	Left-Right	Dorsoventral	Craniocaudal
Residual setup	± 0.38	± 0.32	± 0.21	± 0.61	± 0.48	± 0.52
Intrafraction motion	± 0.28	± 0.16	± 0.18	± 0.42	± 0.25	± 0.32
<b>Total geometrical Error</b>	<b>± 0.87</b>	<b>± 0.80</b>	<b>± 0.82</b>	<b>± 1.03</b>	<b>± 0.89</b>	<b>± 0.94</b>

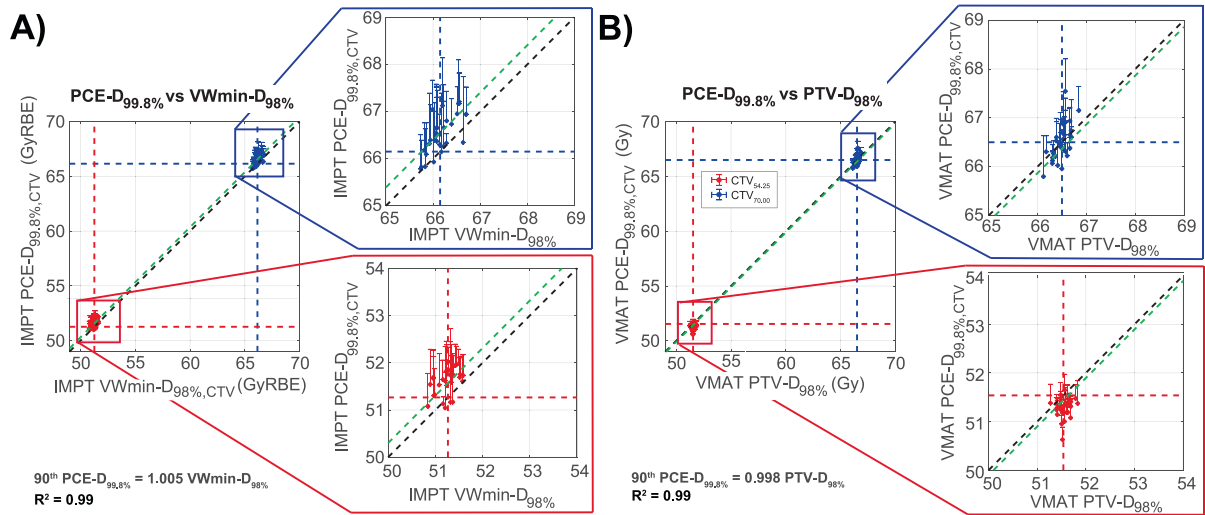


**Fig. 1.** PCE-based robustness evaluation workflow. First, clinical VMAT and IMPT plans were normalized to their prescription doses. This means that VMAT and IMPT treatment plans were normalized to the PTV- $D_{98\%}$  and  $VWmin-D_{98\%,CTV}$ , as found in Figure S1 and S2. Second, 100,000 complete fractionated treatments were simulated. Scenario dose-volume parameter distributions were then calculated according to the simulated treatments. Correlations of the dose-volume metrics for the CTV and OARs between both modalities were done considering percentiles of the scenario distributions.

53.3 (10<sup>th</sup> percentile: 53.1) GyRBE for the  $CTV_{70,00}$  and  $CTV_{54,25}$ , respectively (Fig. 3A). For VMAT, dose values of 66.9 (10<sup>th</sup> percentile: 66.6) Gy and 52.2 (10<sup>th</sup> percentile: 52.1) Gy were found, respectively. Dose inhomogeneity was clinically comparable but slightly higher for IMPT ( $p < 0.05$ ), with a median population HI value of 0.96 (5<sup>th</sup>/95<sup>th</sup> percentiles: 0.96/0.96) and 0.75 (5<sup>th</sup>/95<sup>th</sup> percentiles: 0.75/0.76) for the  $CTV_{70,00}$  and  $CTV_{54,25}$  respectively, while, for VMAT, values of 0.94 (5<sup>th</sup>/95<sup>th</sup> percentiles: 0.94/0.96) and 0.74 (5<sup>th</sup>/95<sup>th</sup> percentiles: 0.73/0.75) were obtained respectively. Higher dose conformity was also achieved for IMPT, specially for the  $CTV_{70,00}$  (Fig. 3C). IMPT resulted on a median population CI of 2.10 (5<sup>th</sup>/95<sup>th</sup> percentiles: 2.03/2.16) for the

$CTV_{54,25}$  and 1.90 (5<sup>th</sup>/95<sup>th</sup> percentiles: 1.85/1.93) for the  $CTV_{70,00}$ , while for VMAT, values of 2.22 (5<sup>th</sup>/95<sup>th</sup> percentiles: 2.20/2.25) and 2.01 (5<sup>th</sup>/95<sup>th</sup> percentiles: 1.98/2.04) were found respectively (Fig. 3D). The slightly higher CTV dose found for IMPT resulted in an increased plan superiority in terms of total  $TCP_T$  compared to VMAT ( $p < 0.05$ ). IMPT showed a higher median population total  $TCP_T$  of 48.1 (5<sup>th</sup>/95<sup>th</sup> percentiles: 47.8/48.7)%, compared to the value of 46.2 (5<sup>th</sup>/95<sup>th</sup> percentiles: 45.8/46.6)% from VMAT (Fig. 3E).

The mean OAR doses involved in the NTCP for xerostomia and dysphagia complications are displayed in Fig. 4A and Fig. 4B respectively. An improved dosimetric benefit for the majority of



**Fig. 2.** Comparison of clinical VWmin-D<sub>98%,CTV</sub> (A) and PTV-D<sub>98%</sub> (B) against the scenario D<sub>99.8%,CTV</sub> distribution. The points correspond to the 10<sup>th</sup> percentile of the sampled treatments with PCE, with an upper error bar representing the median value of the scenario D<sub>99.8%,CTV</sub> distribution. The red and blue dashed lines represent the planning constraints on the CTV<sub>54.25</sub> and CTV<sub>70.00</sub>, while the black and green dashed lines correspond to the identity and regression lines respectively.

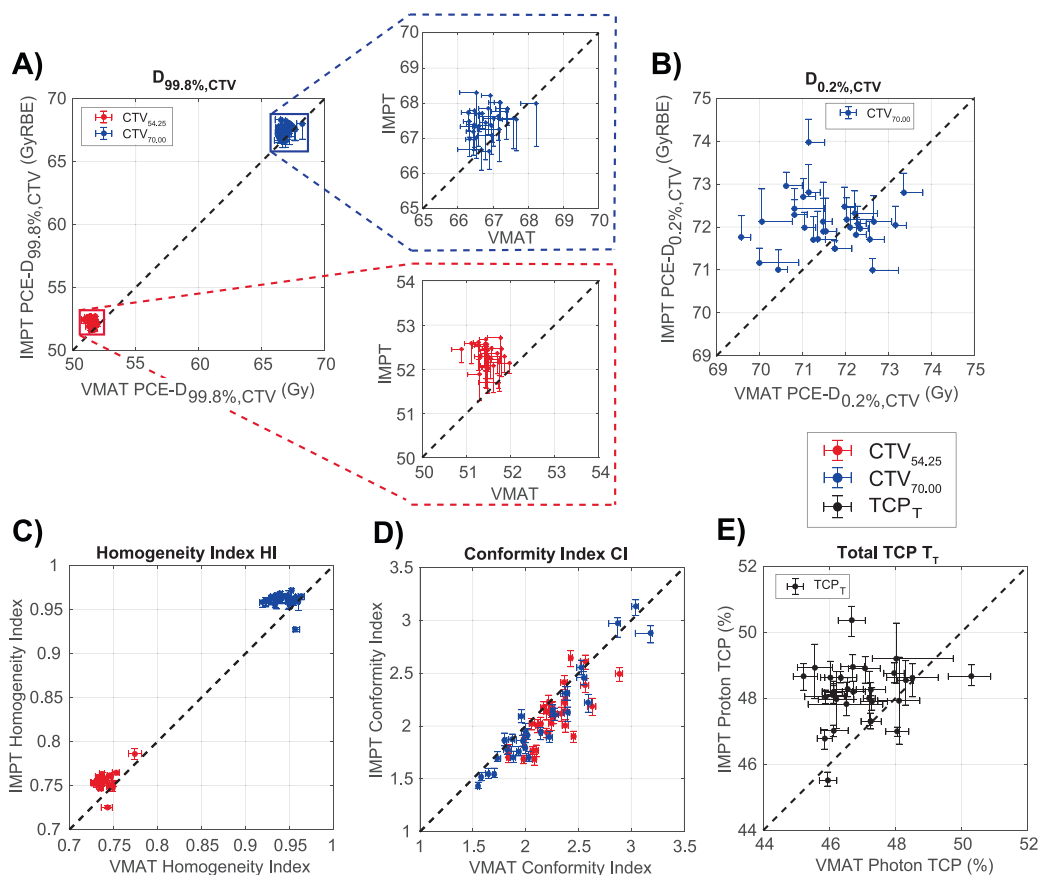
**Table 2**

Comparison of clinical MC 0.1% and PCE-based dose parameters for the CTV and OARs. The values represent the median values of the dose parameters calculated with the clinical dose engine and with PCE. To assess the impact of treatment errors in dose parameters, percentiles of the scenario distributions are also reported (CTV: D<sub>99.8%</sub>, D<sub>98%</sub> V<sub>95%</sub> with the 10<sup>th</sup> percentile, D<sub>2%</sub> and D<sub>0.2%</sub> with the 90<sup>th</sup> percentile; D<sub>mean</sub> of OARs and NTCPs with 5<sup>th</sup> and 95<sup>th</sup> percentiles).

Target Structure Metric	Nominal		VWmin/VWmax	PTV	PCE percentiles	
	IMPT	VMAT			IMPT	VMAT
CTV <sub>70.00</sub>						
D <sub>98%</sub> (Gy/GyRBE)	68.7	67.4	66.3	66.5	68.5 (68.2)	67.4 (67.2)
V <sub>95%</sub> (%)	99.1	98.2	98.5	97.8	99.1 (99.1)	98.2 (98.2)
D <sub>2%</sub> (Gy/GyRBE)	71.4	71.1	73.7	72.4	71.2 (71.6)	70.9 (71.0)
D <sub>99.8%</sub> (Gy/GyRBE)					67.6 (66.8)	66.9 (66.6)
D <sub>0.2%</sub> (Gy/GyRBE)					72.0 (72.6)	71.5 (71.7)
CTV <sub>54.25</sub>						
D <sub>98%</sub> (Gy/GyRBE)	53.4	52.1	51.4	51.5	53.2 (53.1)	52.2 (52.1)
V <sub>95%</sub> (%)	99.1	98.2	98.2	97.9	99.1 (99.1)	98.2 (98.2)
D <sub>99.8%</sub> (Gy/GyRBE)					52.3 (51.9)	51.5 (51.3)
<b>OARs</b>	<b>Nominal</b>		<b>PCE percentiles</b>			
<b>Metric Structure</b>	IMPT	VMAT	IMPT		VMAT	
D <sub>mean</sub> (Gy/GyRBE)						
Parotid Ipsilateral	21.5	27.0	21.6 (19.6–23.6)		27.0 (25.0–29.1)	
Parotid Contralateral	11.3	17.9	11.4 (9.8–13.1)		18.0 (16.4–19.6)	
Submandibular Ipsilateral	55.5	57.1	55.6 (53.7–57.3)		57.5 (56.0–58.2)	
Submandibular Contralateral	35.1	41.6	34.8 (31.8–37.8)		41.1 (33.9–43.0)	
Oral Cavity	27.2	37.7	27.1 (25.3–28.9)		37.7 (36.4–39.0)	
PCM Superior	44.4	50.9	44.5 (43.0–46.0)		51.0 (49.8–52.0)	
PCM Medium	41.0	48.7	41.3 (39.7–43.0)		49.0 (47.7–50.4)	
PCM Inferior	25.3	34.4	25.6 (23.8–27.4)		34.9 (34.0–35.8)	
Average	32.4	38.7	33.1 (3.0–59.0)		39.6 (12.8–60.6)	
<b>NTCPs (%) Toxicity</b>	IMPT	VMAT	IMPT		VMAT	
Xerostomia II	40.1	45.3	40.0 (38.2–41.6)		45.5 (44.3–46.5)	
Dysphagia II	14.9	25.7	15.1 (13.7–16.4)		25.9 (24.7–27.2)	
Xerostomia III	10.9	12.8	10.9 (10.2–11.4)		12.8 (12.4–13.2)	
Dysphagia III	2.4	6.1	2.4 (2.2–2.7)		6.4 (6.0–7.0)	

the OARs was achieved for IMPT compared to VMAT, with an average D<sub>mean</sub> reduction of 6.5 Gy over all patients. In particular, most patients qualified through the dysphagia NTCP models due to the improved oral cavity and PCM dose sparing in IMPT, showing an average NTCP reduction of 10.8 (range: 2.8–16.0) percentage points (%-point) (grade II) and 4.0 (range: 0.6–8.1) %-point (grade III). For grade II and III xerostomia complications respectively, moderate NTCP gains of 5.5 (range: 1.0–11.6) %-point and 1.9 (range: 0.4–4.7) %-point were found.

NTCP spreads for the two treatment modalities showed a reasonable robustness of the MBS approach for patient selection to PT. For grade II xerostomia and dysphagia toxicities, NTCP spreads were significantly (p < 0.05) but non-clinically relevant larger for IMPT, with values of 3.0 (range: 0.5–4.6) %-point and 2.9 (range: 0.2–5.1) %-point while, for VMAT, values of 2.2 (range: 0.3–3.7) %-point and 2.2 (range: 0.6–4.9) %-point were found respectively. For grade III complications, non-significant differences were found between VMAT and IMPT. In this case, NTCP spreads of 1.1 (range: 0.1–2.3) %-point and 0.6 (range: 0.1–4.4) %-point for IMPT and 0.9



**Fig. 3.** Population  $D_{99.8\%,CTV}$  (A),  $D_{0.2\%,CTV}$  (B), Homogeneity (C) and Conformity indexes (D) and  $TCP_T$  (E) comparison between the VMAT and IMPT plans. The points correspond to the median value of the sampled treatments with PCE, with error bars representing the 10<sup>th</sup> percentile for the  $D_{99.8\%,CTV}$ , the 90<sup>th</sup> percentile for the  $D_{0.2\%,CTV}$ , and the 95<sup>th</sup> and 5<sup>th</sup> percentiles for the HI, the CI and the  $TCP_T$ . The black dashed line corresponds to the identity line.

(range: 0.1–2.3) %-point and 0.8 (range: 0.2–4.8) %-point for VMAT were obtained.

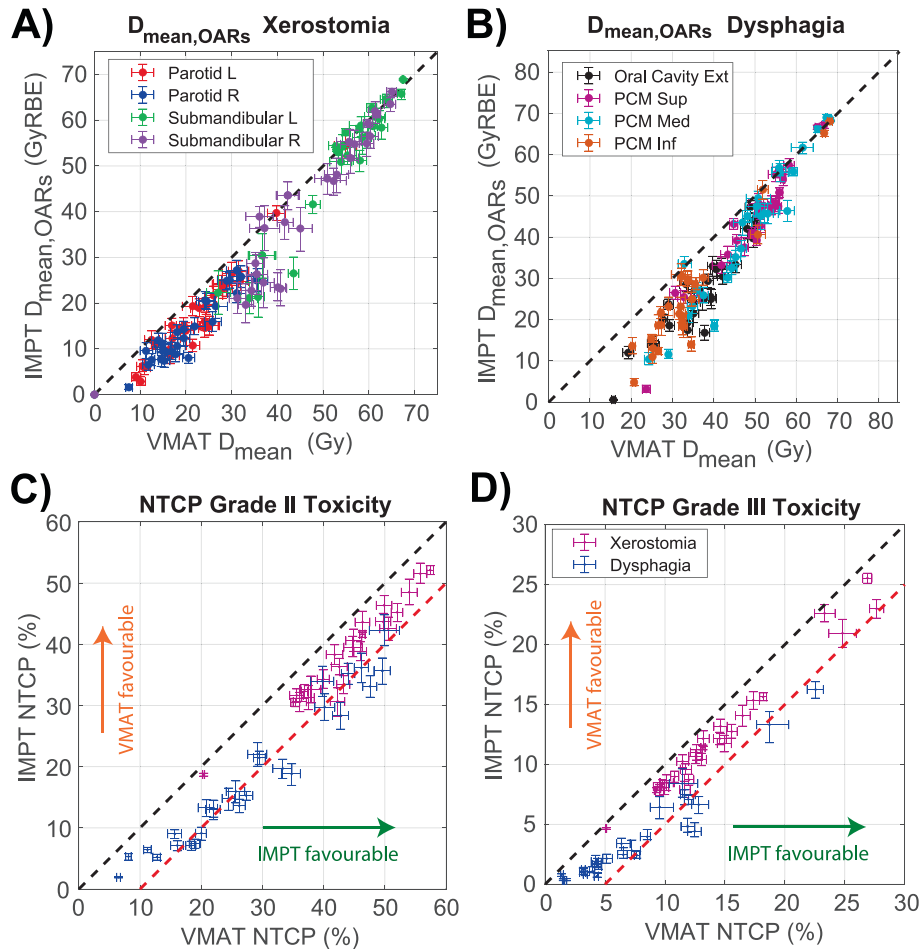
**Discussion**

In this study, plan equivalence between PTV-based VMAT and robustly optimized IMPT was evaluated when the Dutch MBS is used to qualify HNC patients for PT. We compared not only dosimetric CTV and OARs differences but also modelled NTCPs and TCPs outcomes. The impact of geometrical and range errors on the delivered dose was assessed with PCE, in which a fast and accurate robustness evaluation of 100,000 complete fractionated treatments was done per plan. PCE allowed to obtain not only accurate statistics on clinically relevant dosimetric parameters but also probabilistic robustness metrics for the CTV (e.g., scenario  $D_{99.8\%,CTV}$  distributions) and main OARs (e.g., scenario  $D_{mean}$  distributions).

Our main finding is that, in line with the 90% population probability aimed at by van Herk [8], a near-minimum volume  $v = 99.8\%$  probabilistically lead to consistent results compared to clinical plan evaluation metrics used in conventional RT and PT respectively. This volume resulted from the relaxation of the historical clinical goal from a point dose minimum ( $D_{100\%}$ ) to the near-minimum dose ( $D_{98\%}$ ) [22], which was used in the calibration of the DUPROTON protocol [11]. Thus, the 10<sup>th</sup> percentile of the scenario  $D_{99.8\%,CTV}$  distribution showed the best agreement with clinical PTV- $D_{98\%}$  and  $VWmin-D_{98\%,CTV}$  doses. The slight variations found between the clinical metrics and the probabilistic  $D_{99.8\%,CTV}$  values indicate that, even after a calibration of the protocol, differ-

ences between the robustness approaches in protons (robust optimization) and photons (PTV-based evaluation) remains. However, the introduction of probabilistic metrics allows a fair and unbiased robustness comparison and target dose adequacy assessment between both modalities, directly on the CTV and without the necessity of a prior calibration. Furthermore, they also enable to interpret robustness of CTV dose for non-robust patients - patients for whom clinical goals were deliberately not achieved - while this remains a limitation in clinical PTV-based and voxel-wise dose metrics.

Despite the fundamental physical differences between photons and protons, the planning comparison between VMAT and IMPT leads to reasonably consistent and robust results in terms of CTV dose, TCP and NTCP. Even with the prescription iso-dose correction of 1% from the DUPROTON calibration [11], IMPT plans resulted in systematically higher nominal doses to the CTV, which also resulted in a significant plan superiority in terms of  $TCP_T$ . This slight increase in  $TCP_T$  for IMPT comes mainly from: First (i), the impact of MC noise, which was not considered in the DUPROTON calibration. As target dose metrics e.g.,  $D_{99.8\%,CTV}$  and  $D_{0.2\%,CTV}$ , are assessed in the tail of the dose-volume histogram distributions, larger underestimations and overestimations of these metrics are found, when the magnitude of the MC noise is increased [SM, Section S6]. This leads to an inherent increase of the robustness in the IMPT plans, which can be corrected with an additional calibration in the clinical voxel-wise values. Second (ii), the normalization of the treatment plans. This could be considered as a limitation of the study, in which variations of target near minimum dose values from the clinical goals were removed. As such, results do not show



**Fig. 4.** Population  $D_{mean}$  for the main relevant OARs (A and B) and NTCP for grade II (C) and grade III (D) xerostomia and dysphagia comparison between the VMAT and IMPT plans. The points correspond to the median value of the sampled treatments with PCE, with error bars representing the 95<sup>th</sup> and 5<sup>th</sup> percentiles. The dashed lines correspond to the identity line (black) and to the threshold to qualify for IMPT (red).

the actual clinical differences between VMAT and IMPT plans in the studied patient cohort, but it shows differences that can be expected when plans are created such that the clinical goals for target near minimum dose are exactly met. Due to the superior target near minimum dose of VMAT plans compared to IMPT in the cohort, the normalization resulted in a reduced  $D_{mean}$  dose in GTVs and both CTVs for VMAT compared to IMPT, and consequently inferior TCP<sub>T</sub> values (Fig. S4).

The moderate impact of geometrical and range errors on NTCPs (NTCP spreads < 3%) showed that the MBS approach is robust. It showed that the nominal plan can be generally used as a good estimator to refer patients for PT. Furthermore, the probabilistic assessment of NTCPs could also potentially improve NTCP-based protocols for PT patient selection. The inclusion of bandwidths to the  $\Delta$ NTCP thresholds of the protocol could particularly benefit some VMAT borderline patients, who had a non-negligible fraction of simulated treatments qualifying for IMPT.

PCE has demonstrated to be a powerful robustness evaluation tool compared to other robustness evaluation methods [11,26], as it enables a comprehensive robustness analysis assuming a large number of sampled dose distributions. However, to convert the technical superiority into clinical success might not be a trivial task. Clinically introduced methods to present valuable and intuitive spatial information, such as the slice-by-slice visualization of under- and over-dose areas within the CTV given by traditional PTV-based and voxel-wise evaluations, need to be established for PCE or any other probabilistic plan evaluation approach. A first step

could be to compute with PCE voxel-wise doses, in which VWmin and VWmax dose distributions are generated from the 90% of the PCE sampled doses, excluding the most extreme scenarios [SM, section S7]. The high correlation shows that such a method would not lead to a systematic bias when used to replace PTV/VWmin evaluations. PCE could also generate probability dose maps – per voxel probability values to have a dose below/above a certain dose constraint – to spatially locate areas with a high under/overdose probability.

Our study is limited to isocentric errors, which were modelled as rigid shifts and SPP errors for IMPT. As our goal was to probabilistically compare initial plans, anatomical variations were not included, although the same methodology could be applied at any time during the course of treatment when diagnostic quality verification CTs are acquired. The combination of PCE-based robustness evaluations with online or offline adaptive strategies could assess the suitability of a plan with the patient anatomy of the day, potentially improving clinical decisions for plan adaptations [27,28]. Residual patient rotations were not considered in this study.

As PCE is an analytical approximation of the dose engine, a proper validation of the model is required. The PCE model parameters were selected such that, for sufficiently large errors up to 4.5 mm, PCE dose distributions agrees within 1% with the clinical doses. The large number of spots per energy layer used in the clinical IMPT plans lead to highly heterogeneous dose distributions, which compromised PCE accuracy. Thus, the accuracy of the PCE



models for IMPT was increased using a low MC uncertainty of 0.1%. PCE accuracy could also be improved by increasing (i) the order of the polynomial basis vectors and (ii) the amount of clinical doses required during the PCE construction, which however also increases the computational time to construct the model. For more complex treatment sites, the parametrization of uncertainties might also be difficult or even impossible e.g., breathing-related intra-fraction errors, in which a combination of PCE with other robustness approaches could be used [29].

## Conclusion

Despite the differences between photon and proton planning, the comparison between PTV-based VMAT and robustly optimized IMPT after the proton plan calibration is consistent, with a slightly higher dose to the CTV for IMPT. The MBS procedure is moderately impacted by the effect of geometrical and range errors on NTCP, showing that the nominal plan is a good estimator to refer patients for PT. Finally, to the best of our knowledge, this study is the first to present comprehensive probabilistic plan evaluation metrics on the CTV to assess robustness between proton and photon treatments, which are consistent with clinical PTV-based and voxel-wise metrics and could be used in the future for probabilistic planning.

## Conflict of interest

This work is supported by KWF Kanker Bestrijding (project number 11711). The department of radiotherapy of the Erasmus MC Cancer Institute has also research collaborations with Elekta AB, Stockholm, Sweden and Accuray Inc., Sunnyvale, USA.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was funded by the KWF Dutch cancer society (KWF project number 11711). The authors would like to thank Mirko Unipan for the insightful discussions, Dirk Wagenaar for the help in the TCP modelling and Rob Louwe for the analysis of machine quality assurance data. Zoltán Perkö would like to thank the support of the NWO VENI grant ALLEGRO (016.Veni.198.055) during the time of this study.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.radonc.2023.109729>.

## References

- [1] van der Veen J, Nuyts S. Can intensity-modulated-radiotherapy reduce toxicity in head and neck squamous cell carcinoma?, 2017.
- [2] McKeever MR, Sio TT, Brandon Gunn G, et al. Reduced acute toxicity and improved efficacy from intensity-modulated proton therapy (IMPT) for the management of head and neck cancer. *Chin Clin Oncol* 2016.

- [3] Langendijk JA, Hoebbers FJ, De Jong MA, et al. National protocol for model-based selection for proton therapy in head and neck cancer. *Int J Part Ther* 2021.
- [4] Tamba M, Steenbakkers RJ, van der Laan HP, et al. First experience with model-based selection of head and neck cancer patients for proton therapy. *Radiother Oncol* 2020.
- [5] Landelijk Platform Protontherapie, Protontherapie LP. Landelijk Indicatie Protocol Protonen Therapie Hoofd-halstumoren. 2019.
- [6] Lomax AJ. Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: The potential effects of inter-fraction and inter-field motions. *Phys Med Biol* 2008.
- [7] Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations, 2012.
- [8] Van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol* 2004.
- [9] Fredriksson A, Forsgren A, Hardemark B. Minimax optimization for handling range and setup uncertainties in proton therapy. *Med Phys* 2011.
- [10] Fredriksson A, Forsgren A, Hardemark B. Minimax optimization for handling range and setup uncertainties in proton therapy. *Med Phys* 2011;38:1672–84.
- [11] Korevaar EW, Habraken SJM, Scandurra D, et al. Practical robustness evaluation in radiotherapy—A photon and proton-proof alternative to PTV-based plan evaluation. *Radiother Oncol* 2019;141:267–74.
- [12] Perko Z, Van Der Voort SR, Van De Water S, Hartman CM, Hoogeman M, Lathouwers D. Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion. *Phys Med Biol* 2016;61:4646–64.
- [13] Rojo-Santiago J, Habraken SJ, Lathouwers D, Mendez Romero A, Perko Z, Hoogeman MS. Accurate assessment of a Dutch practical robustness evaluation protocol in clinical PT with pencil beam scanning for neurological tumors. *Radiother Oncol* 2021.
- [14] Perkö Z. Open source generalized Polynomial Chaos Expansion (openGPC) Toolbox. <https://gitlab.com/zperko/opengpc>.
- [15] Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol* 2015.
- [16] Kierkels RG, Korevaar EW, Steenbakkers RJ, et al. Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans. *Radiother Oncol* 2014.
- [17] Wagenaar D, Kierkels R, van der Schaaf A, et al. Robustness evaluation using dose accumulation in head and neck IMPT. *Int J Radiat Oncol Biol Phys* 2019.
- [18] Lomax AJ. Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: The potential effects of calculational uncertainties. *Phys Med Biol* 2008.
- [19] Meijers A, Free J, Wagenaar D, et al. Validation of the proton range accuracy and optimization of CT calibration curves utilizing range probing. *Phys Med Biol* 2020.
- [20] Meijers A, Seller Oria C, Free J, Langendijk JA, Knopf AC, Both S. Technical note: First report on an in vivo range probing quality control procedure for scanned proton beam therapy in head and neck cancer patients. *Med Phys* 2021.
- [21] Taasti VT, Muren LP, Jensen K, et al. Comparison of single and dual energy CT for stopping power determination in proton therapy of head and neck cancer. *Phys Imaging Radiat Oncol* 2018.
- [22] Menzel HG. The international commission on radiation units and measurements. *J ICRU* 2010.
- [23] Lühr A, Lock S, Jakobi A, et al. Modeling tumor control probability for spatially inhomogeneous risk of failure based on clinical outcome data. *Z Med Phys* 2017.
- [24] Due AK, Vogelius IR, Aznar MC, et al. Recurrences after intensity modulated radiotherapy for head and neck squamous cell carcinoma more likely to originate from regions with high baseline [18F]-FDG uptake. *Radiother Oncol* 2014.
- [25] Jakobi A, Lühr A, Stutzer K, et al. Increase in tumor control and normal tissue complication probabilities in advanced head-and-neck cancer for dose-escalated intensity-modulated photon and proton therapy. *Front Oncol* 2015.
- [26] Sterpin E, Rivas ST, Van Den Heuvel F, George B, Lee JA, Souris K. Development of robustness evaluation strategies for enabling statistically consistent reporting. *Phys Med Biol* 2021.
- [27] Oud M, Breedveld S, Giżyńska M, et al. An online adaptive plan library approach for intensity modulated proton therapy for head and neck cancer. *Radiother Oncol* 2022;176:68–75.
- [28] Borderias-Villarreal E, Taasti V, Van Elmpt W, Teruel-Rivas S, Geets X, Sterpin E. Evaluation of the clinical value of automatic online dose restoration for adaptive proton therapy of head and neck cancer. *Radiother Oncol* 2022;170:190–7.
- [29] Pastor-Serrano O, Habraken S, Lathouwers D, Hoogeman M, Schaart D, Perko Z. How should we model and evaluate breathing interplay effects in IMPT? *Phys Med Biol* 2021.