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

Publication date
2023

Document Version
Final published version

Published in
Radiotherapy and Oncology

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.
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,a,b,1,2,* Erik Korevaar, Zoltán Perkó, Stefan Both, Steven J.M. Habraken,a,b, Mischa S. Hoogeman,a,b

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

A B S T R A C T

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 \(V\min-\text{D}_{98,\text{CTV}}\) 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 \(D_{98,\text{CTV}}\) distribution. Most patients qualified for IMPT through the dysphagia grade II model, for which an average NTCP gain of 10.5 percentages 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/).
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 (CTV1000) and 54.25 Gy to the elective lymph nodes (CTV423), both delivered in 35 fractions. For VMAT, dose was prescribed to the PTV: PTV-D$_{95\%}$ ≥ $D_{\text{pres}}$, while, for IMPT, dose was prescribed to the voxel-wise near-minimum (VWMmin) CTV dose: VWMmin-D$_{\text{pres},\text{CTV}}$ ≥ 94% $D_{\text{pres}}$ [11]. To avoid hot spots in the target, an additional constraint was included for the PTV near-maximum dose: PTV-D$_{95\%}$ ≤ 107% $D_{\text{pres}}$ for VMAT and on the voxel-wise near-maximum (VWMmax) CTV dose: VWMmax-D$_{1\text{cc},\text{CTV}}$ ≤ 78 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 computerized tomography (SPECT) scan, with 0.98 × 0.98 × 2 mm$^3$ CT resolution. The median CTV was 13.49 cc (range 1.37–117.05 cc). Main relevant OARs i.a., 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° and 320°) and 2 posterior (160° and 200°) 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 VWMmin and VWMmax 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 VWMmin-D$_{\text{pres},\text{CTV}}$ (G, IMPT) and PTV-D$_{95\%}$ (G$_{\text{VMAT}}$) 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, 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 $\vec{s} = (\vec{s}_x, \vec{s}_y, \vec{s}_z)$ and a relative range error $\rho$ as the series expansion $D_i(\vec{s}, \rho) = \sum_{\vec{s}, \rho} a_{\vec{s}, \rho} \Psi_k(\vec{s}, \rho)$, with expansion coefficients $\{a_{\vec{s}, \rho}\}$ and multi-dimensional Hermite polynomials $\Psi_k(\vec{s}, \rho)$ [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 mm and $\sigma$ = 1.00 mm for the systematic and random geometrical errors (1 SD) were assumed, consistent with a $M = 2.5 \Sigma + 0.7 \sigma = 3$ mm
A near-minimum dose (PTV-D 98%) was found for IMPT, resulting in a median geneity index (HI) value, defined as the ratio between \( D_{95}\) and \( D_{98}\) for the CTV70.00 and CTV54.25.

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, CTV70.00 and CTV54.25) with different response to dose [23–25]. TCP scenario distributions for the GTV, disjoint CTV70.00 and disjoint CTV54.25 were calculated from the dose-volume histograms of the treatments sampled in each of the regions. The total TCP \( (TCP_{T}) \) was calculated as: \( TCP_{T} = TCP_{GTV} \times TCP_{CTV70} \times TCP_{CTV54} \). 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 \( \Delta NTCP = NTCP_{VMAT} - NTCP_{IMPT} \) 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 Wilcoxon signed rank test \( (p < 0.05) \).

**Results**

Based on the calibration for the probabilistic \( D_{v,CTV} \) metric [SM, Section S4], the 10th percentile of \( D_{v} = 99.8\% (D_{99.8\%CTV}) \) correlated best with clinical VWmin-\( D_{98\%CTV} \) for IMPT and PTV-\( D_{98\%} \) for VMAT values. As depicted in Fig. 2, the 10th percentile of the scenario \( D_{99.8\%CTV} \) distributions for both CTVs (CTV70.00 and CTV54.25) 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-\( D_{98\%CTV} \) (\( R^{2} > 0.62 \)) and PTV-\( D_{98\%} \) (\( R^{2} > 0.94 \)). 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 \( D_{98\%CTV} \) of 67.6 (10th percentile: 66.8) GyRBE and of...
53.3 (10th percentile: 53.1) GyRBE for the CTV70.00 and CTV54.25, respectively (Fig. 3A). For VMAT, dose values of 66.9 (10th percentile: 66.6) Gy and 52.2 (10th 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 (5th/95th percentiles: 0.96/0.96) and 0.75 (5th/95th percentiles: 0.75/0.76) for the CTV70.00 and CTV54.25, respectively, while, for VMAT, values of 0.94 (5th/95th percentiles: 0.94/0.96) and 0.74 (5th/95th percentiles: 0.73/0.75) were obtained respectively. Higher dose conformality was also achieved for IMPT, especially for the CTV70.00 (Fig. 3C). IMPT resulted on a median population CI of 2.10 (5th/95th percentiles: 2.03/2.16) for the CTV54.25 and 1.90 (5th/95th percentiles: 1.85/1.93) for the CTV70.00, while for VMAT, values of 2.22 (5th/95th percentiles: 2.20/2.25) and 2.01 (5th/95th 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 compared to VMAT (p < 0.05). IMPT showed a higher median population total TCP of 48.1 (5th/95th percentiles: 47.8/48.7)% compared to the value of 46.2 (5th/95th 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 patients was observed for IMPT compared to VMAT.
the OARs was achieved for IMPT compared to VMAT, with an average $D_{\text{mean}}$ 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.

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: $D_{\text{99.8%}}$, $D_{\text{98%}}$ $V_{\text{95%}}$ with the 10th percentile, $D_{\text{2%}}$ and $D_{\text{0.2%}}$ with the 90th percentile; $D_{\text{mean}}$ of OARs and NTCPs with 5th and 95th percentiles).

<table>
<thead>
<tr>
<th>Target</th>
<th>Structure Metric</th>
<th>Nominal PCE percentiles</th>
<th>PT</th>
<th>PCE percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IMPT</td>
<td>VMAT</td>
<td>IMPT</td>
<td>VMAT</td>
</tr>
<tr>
<td>CTV54.25</td>
<td>$D_{\text{mean}}$ (Gy/GyRBE)</td>
<td>68.7</td>
<td>67.4</td>
<td>66.3</td>
</tr>
<tr>
<td></td>
<td>$V_{\text{95%}}$ (%)</td>
<td>99.1</td>
<td>98.2</td>
<td>98.5</td>
</tr>
<tr>
<td></td>
<td>$D_{\text{99.8%}}$ (Gy/GyRBE)</td>
<td>71.4</td>
<td>71.1</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td>$D_{\text{98%}}$ (Gy/GyRBE)</td>
<td>53.4</td>
<td>52.1</td>
<td>51.4</td>
</tr>
<tr>
<td></td>
<td>$V_{\text{95%}}$ (%)</td>
<td>99.1</td>
<td>98.2</td>
<td>98.2</td>
</tr>
<tr>
<td>OARs</td>
<td>$D_{\text{mean}}$ (Gy/GyRBE)</td>
<td>53.4</td>
<td>52.1</td>
<td>51.4</td>
</tr>
<tr>
<td></td>
<td>Nominal PCE percentiles</td>
<td>IMPT</td>
<td>VMAT</td>
<td>IMPT</td>
</tr>
<tr>
<td></td>
<td>Parotid Ipsilateral</td>
<td>21.5</td>
<td>27.0</td>
<td>21.6 (19.6–23.6)</td>
</tr>
<tr>
<td></td>
<td>Parotid Contralateral</td>
<td>11.3</td>
<td>17.9</td>
<td>11.4 (9.8–13.1)</td>
</tr>
<tr>
<td></td>
<td>Submandibular Ipsilateral</td>
<td>55.5</td>
<td>57.1</td>
<td>55.6 (53.7–57.2)</td>
</tr>
<tr>
<td></td>
<td>Submandibular Contralateral</td>
<td>35.1</td>
<td>41.6</td>
<td>34.8 (31.8–37.8)</td>
</tr>
<tr>
<td></td>
<td>Oral Cavity</td>
<td>27.2</td>
<td>37.7</td>
<td>27.1 (25.3–28.9)</td>
</tr>
<tr>
<td></td>
<td>PCM Superior</td>
<td>44.4</td>
<td>50.9</td>
<td>44.5 (43.0–46.0)</td>
</tr>
<tr>
<td></td>
<td>PCM Medium</td>
<td>41.0</td>
<td>48.7</td>
<td>41.3 (39.7–43.0)</td>
</tr>
<tr>
<td></td>
<td>PCM Inferior</td>
<td>25.3</td>
<td>34.4</td>
<td>25.6 (23.8–27.4)</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>32.4</td>
<td>38.7</td>
<td>33.1 (3.0–59.0)</td>
</tr>
<tr>
<td>NTCPs (%)</td>
<td>Toxicity</td>
<td>IMPT</td>
<td>VMAT</td>
<td>IMPT</td>
</tr>
<tr>
<td>Xerostomia II</td>
<td>40.1</td>
<td>45.3</td>
<td>40.0 (38.2–41.6)</td>
<td>45.5 (44.3–46.5)</td>
</tr>
<tr>
<td>Dysphagia II</td>
<td>14.9</td>
<td>25.7</td>
<td>15.1 (13.7–16.4)</td>
<td>25.9 (24.7–27.2)</td>
</tr>
<tr>
<td>Dysphagia III</td>
<td>10.9</td>
<td>12.8</td>
<td>10.9 (10.2–11.4)</td>
<td>12.5 (12.4–13.2)</td>
</tr>
<tr>
<td>Dysphagia III</td>
<td>2.4</td>
<td>6.1</td>
<td>2.4 (2.2–2.7)</td>
<td>6.4 (6.0–7.0)</td>
</tr>
</tbody>
</table>
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, 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]. IMPT plans resulted in systematically higher nominal doses to the CTV, which also resulted in a significant plan superiority in terms of TCP. This slight increase in TCP 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. 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. 3. Population D 99.8% CTV (A), D 0.2% CTV (B), Homogeneity (C) and Conformity indexes (D) and TCP (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 10th percentile for the D 99.8% CTV, the 90th percentile for the D 0.2% CTV, and the 95th and 5th percentiles for the HI, the CI and the TCP. The black dashed line corresponds to the identity line.
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 $D_{\text{mean}}$ dose in GTVs and both CTVs for VMAT compared to IMPT, and consequently inferior TCP 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 ATCP 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

![Fig. 4. Population $D_{\text{mean}}$ 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 95th and 5th percentiles. The dashed lines correspond to the identity line (black) and to the threshold to qualify for IMPT (red).]
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 Perkó 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

[22] 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.