A probabilistic evaluation of the Dutch robustness and model-based selection protocols for Head-and-Neck IMPT

A multi-institutional study

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

DOI
10.1016/j.radonc.2024.110441

Publication date
2024

Document Version
Final published version

Published in
Radiotherapy and Oncology

Radiotherapy and Oncology 199 (2024) 110441

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

A probabilistic evaluation of the Dutch robustness and model-based selection protocols for Head-and-Neck IMPT: A multi-institutional study

Jesús Rojo-Santiago [a,b,*], Steven J.M. Habraken [a,b], Mirko Unipan [c], Stefan Both [d], Geert Bosmans [c], Zoltán Perkó [e], Erik Korevaar [d], Mischa S. Hoogeman [a,b]

[a] Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, the Netherlands
[b] HollandPTC, Delft, the Netherlands
[c] GROW School for Oncology, Maastricht University Medical Center, Department of Radiation Oncology (Maastro), Maastricht, the Netherlands
[d] Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[e] Delft University of Technology, Department of Radiation Science and Technology, Delft, the Netherlands

ARTICLE INFO

Keywords:
Probabilistic planning
Robustness evaluation
Robustness analysis
Robust planning
Model-based selection
Proton therapy
Polynomial Chaos Expansion
Head and Neck

ABSTRACT

Background and purpose: In the Netherlands, 2 protocols have been standardized for PT among the 3 proton centers: a robustness evaluation (RE) to ensure adequate CTV dose and a model-based selection (MBS) approach for IMPT patient-selection. This multi-institutional study investigates (i) inter-patient and inter-center variation of target dose from the RE protocol and (ii) the robustness of the MBS protocol against treatment errors for a cohort of head-and-neck cancer (HNC) patients treated in the 3 Dutch proton centers.

Materials and methods: Clinical treatment plans of 100 HNC patients were evaluated. Polynomial Chaos Expansion (PCE) was used to perform a comprehensive robustness evaluation per plan, enabling the probabilistic evaluation of 100,000 complete fractionated treatments. PCE allowed to derive scenario distributions of clinically relevant dosimetric parameters to assess CTV dose ($D_{99.8\%}/D_{0.2\%}$, based on a prior photon plan calibration) and tumour control probabilities (TCP) as well as the evaluation of the dose to OARs and normal tissue complication probabilities (NTCP) per center.

Results: For the $CTV_{70.00}$, doses from the RE protocol were consistent with the clinical plan evaluation metrics used in the 3 centers. For the $CTV_{54.25}$, $D_{99.8\%}$ were consistent with the clinical plan evaluation metrics at center 1 and 2 while, for center 3, a reduction of 1 GyRBE was found on average. This difference did not impact modelled TCP at center 3. Differences between expected and nominal NTCP were below 0.3 percentage point for most patients.

Conclusion: The standardization of the RE and MBS protocol lead to comparable results in terms of TCP and the NTCPs. Still, significant inter-patient and inter-center variation in dosimetric parameters remained due to clinical practice differences at each institution. The MBS approach is a robust protocol to qualify patients for PT.

Around 500,000 patients are annually diagnosed with Head-and-Neck Cancer (HNC) worldwide, resulting in approximately 270.000 deaths per year [1]. In developed countries, a standard-of-care treatment for HNC patients is the combination of radiation therapy (RT), as intensity-modulated radiation therapy (IMRT), with chemotherapy [2]. To reduce toxicity, intensity-modulated proton therapy (IMPT) has also been clinically introduced as an alternative RT technique. IMPT allows to achieve *iso*-effective target dose with an improvement in healthy tissue sparing compared to conventional RT [3,4]. The reduction of dose to healthy tissue has been demonstrated to improve the quality of life specially for HNC patients [5]. For these patients, an adequate balance between tumour control probability and organs-at-risk (OARs) toxicity is challenging, while a dose reduction could minimize the development of long-term side effects e.g., xerostomia (dry mouth) and dysphagia (swallowing problems) [6–8]. However, IMPT dose delivery: (1) is more expensive and capacity is limited compared to conventional RT [9] and (2) is subject to stopping-power prediction (SPP: range) errors and more sensitive to beam- and patient-alignment (geometrical) inaccuracies [10,11].

In the Netherlands, two protocols have been standardized clinically

* Corresponding author at: P.O. Box 2040, Doctor Molewaterplein 40, 3015GD, Rotterdam, The Netherlands.
E-mail address: j.rojosantiago@erasmusmc.nl (J. Rojo-Santiago).

https://doi.org/10.1016/j.radonc.2024.110441
Received 21 March 2024; Received in revised form 12 July 2024; Accepted 15 July 2024
Available online 26 July 2024

among the three proton centers: First, a robustness evaluation (RE) protocol that, in combination with robust optimization, ensures adequate target dose against treatment uncertainties. In this RE protocol the collaboration of the Dutch proton therapy centers (DUPROTON) has established a dose prescription for the clinical target volume (CTV) to the voxel-wise minimum (VWmin) dose of a standardized set of 28 evaluation scenarios, assuming different combinations of rigid geometrical and range errors [12]. Second, a model-based selection (MBS) protocol was established to select patients for IMPT. The MBS approach consists on an IMRT-IMPT plan comparison to select patient for PT in terms of target and OARs dose and normal tissue complication probabilities (NTCPs) [13–15]. The two protocols have been put in clinical practice before the first patients were treated with IMPT at each center. However, they have some limitations: (i) the calibration of the MBS and RE protocols was photon-based; (ii) the MBS protocol, as derived from clinical NTCP outcome data, depend on delivered dose rather than nominal planned dose; (iii) the RE protocol is limited to a fixed small sample of error scenarios, while actual errors follow continuous distributions and vary from day-to-day. In earlier work [16–18], we evaluated and determined metrics to probabilistically assess the consistency of these protocols at one of the centers. However, their consistency on a multi-institutional level has never been performed. The impact of differences in dose prescription and in treatment planning due to clinical practice variations and adaptations to each treatment delivery system per center (IBA: Louvain-La-Neuve Belgium, Varian: Palo Alto United States of America and Mevion: Littleton United States of America), is unclear.

The purpose of this study was to assess the performance of these protocols for cohorts of head-and-neck cancer (HNC) patients treated in the three Dutch proton centers. Polynomial Chaos Expansion (PCE) was applied to fast and accurately model the dependence of the delivered dose on geometrical and range errors [19,20]. PCE enabled the simulation of 100,000 complete fractionated treatments per plan, to derive probabilistic distributions of dose-volume histogram (DVH) parameters for the primary and elective CTVs and the NTCPs for xerostomia and dysphagia. Subsequently, in a multi-institutional setting, we assessed: (i) inter-patient variation of the CTV dose and tumour control probability (TCP) against geometrical and range errors; and (ii) robustness of the NTCP-based MBS protocol against geometrical and range errors as a referral tool for patients for IMPT.

Table 1
Results of clinical voxel-wise and PCE-based dose parameters for the CTV between the 3 institutions. The table includes the results for the SR=3 mm and 5 mm plans. Clinical VWmin/VWmax $D_{98\%,CTV}/D_{2\%,CTV}$ values are compared with PCE-based $D_{99.8\%}/D_{0.2\%,}$ metrics evaluated at a 90 % population probability, in line with van Herk. Each of the values reports the patient average value and the inter-patient variation as 5th/95th percentiles in brackets. PCE-based $D_{98\%}$, $D_{2\%}$, $V_{95\%}$ values are also reported.

SR	**Target**	**Dose Metric**	**Institution**		
3 mm (60/100)	**Structure**	**Clinical**	**Center 1 (20/100)**	**Center 2 (20/100)**	**Center 3 (20/100)**
	$CTV_{70.00}$				
		Prescribed			
		VWmin-$D_{98\%}$ (GyRBE)	0.94 D_{pres} (65.8)	0.95 D_{pres} (66.5)	0.94 D_{pres} (65.8)
		VWmin-$D_{98\%}$ (GyRBE)	66.8 (66.2–67.7)	67.2 (66.7–68.2)	65.6 (65.1–66.3)
		Prescribed			
		VWmax-$D_{2\%}$ (GyRBE)	1.07 D_{pres} (74.9)	1.07 D_{pres} (74.9)	1.09 D_{pres} (76.3)
		VWmax-$D_{2\%}$ (GyRBE)	73.7 (72.7–75.0)	73.7 (72.9–74.3)	75.3 (74.4–76.0)
		PCE Metric: 90 % population probability	**Center 1**	**Center 2**	**Center 3**
		$D_{99.8\%}$ (GyRBE)	67.3 (66.9–68.4)	67.4 (66.5–68.4)	65.6 (64.0–66.9)
		$D_{98\%}$ (GyRBE)	68.8 (68.4–69.2)	69.0 (68.7–69.5)	68.0 (67.3–68.6)
		$V_{95\%}$ (%)	99.93 (99.88–99.98)	99.93 (99.81–99.99)	99.52 (98.86–99.89)
		$D_{0.2\%}$ (GyRBE)	73.1 (72.3–74.1)	73.7 (72.9–74.6)	75.1 (74.2–75.8)
		$D_{2\%}$ (GyRBE)	72.1 (71.6–72.8)	72.7 (72.0–73.3)	74.1 (73.7–74.7)
	$CTV_{54.25}$				
		Prescribed	0.94 D_{pres} (51.0)	0.95 D_{pres} (51.5)	0.93 D_{pres} (50.5)
		VWmin-$D_{98\%}$ (GyRBE)			
		VWmin-$D_{98\%}$ (GyRBE)	51.7 (51.3–52.2)	52.2 (51.2–52.7)	50.3 (49.6–52.2)
		PCE Metric: 90 % population probability	**Center 1**	**Center 2**	**Center 3**
		$D_{99.8\%}$ (GyRBE)	52.2 (51.7–52.6)	52.3 (51.2–52.8)	49.3 (47.7–51.4)
		$D_{98\%}$ (GyRBE)	53.4 (53.2–53.8)	53.8 (53.5–54.1)	53.2 (52.2–54.8)
		$V_{95\%}$ (%)	99.91 (99.83–99.96)	99.89 (99.75–99.96)	99.30 (98.78–99.75)
5 mm (40/100)	**Structure**	**Clinical**	**Center 1 (20/100)**	**Center 2 (20/100)**	
	$CTV_{70.00}$				
		Prescribed	0.94 D_{pres} (65.8)	0.95 D_{pres} (66.5)	
		VWmin-$D_{98\%}$ (GyRBE)			
		VWmin-$D_{98\%}$ (GyRBE)	67.0 (65.3–68.2)	67.0 (65.7–68.0)	
		Prescribed			
		VWmax-$D_{2\%}$ (GyRBE)	1.07 D_{pres} (74.9)	1.07 D_{pres} (74.9)	
		VWmax-$D_{2\%}$ (GyRBE)	74.6 (73.2–75.7)	73.7 (72.8–74.8)	
		PCE Metric: 90 % population probability	**Center 1**	**Center 2**	
		$D_{99.8\%}$ (GyRBE)	67.6 (65.7–68.5)	67.6 (66.3–68.5)	
		$D_{98\%}$ (GyRBE)	69.0 (68.4–69.5)	69.1 (68.4–69.6)	
		$V_{95\%}$ (%)	99.93 (99.64–99.99)	99.92 (99.71–99.99)	
		$D_{0.2\%}$ (GyRBE)	73.9 (72.8–74.9)	73.3 (72.5–74.4)	
		$D_{2\%}$ (GyRBE)	72.2 (71.8–73.3)	72.4 (71.7–73.2)	
	$CTV_{54.25}$				
		Prescribed			
		VWmin-$D_{98\%}$ (GyRBE)	0.94 D_{pres} (51.0)	0.94 D_{pres} (51.5)	
		VWmin-$D_{98\%}$ (GyRBE)	52.0 (51.0–52.8)	51.9 (50.6–52.8)	
		PCE Metric: 90 % population probability	**Center 1**	**Center 2**	
		$D_{99.8\%}$ (GyRBE)	52.6 (51.9–53.1)	52.3 (50.4–53.2)	
		$D_{98\%}$ (GyRBE)	53.6 (53.2–54.2)	54.0 (53.5–54.5)	
		$V_{95\%}$ (%)	99.94 (99.84–99.99)	99.88 (99.64–99.98)	

Methods and materials

Patient data

A total of 100 HNC patients with oral cavity, pharynx and larynx tumours, treated with IMPT at the Holland Proton Therapy Center (HollandPTC), at the UMC Groningen Proton Therapy Center (UMCG) and at the Maastro Proton Therapy Center (MAASTRO) centers, were included. Patients at HollandPTC were treated from January 2019 to January 2022, while in UMCG were treated between February 2018 to July 2021 and at Maastro patients were treated between April 2018 to September 2021. The analysis of the patient cohorts was anonymized per center, resulting in 40/100 plans for center 1, 40/100 for center 2 and 20/100 plans for center 3. Patients were referred for IMPT through the MBS protocol, nationally known as the Dutch National Indication Protocols for PT (NIPP) (v2.1 and v2.2) [12,13]. Patients were treated to 70 GyRBE for the primary CTV ($CTV_{70.00}$) and 54.25 GyRBE to the elective lymph nodes ($CTV_{54.25}$), both delivered in 35 fractions assuming a constant relative biological effectiveness (RBE) of 1.1. According to the RE protocol [12], dose was prescribed (D_{pres}) to the voxel-wise minimum dose (VWmin) of 28 scenarios: VWmin-$D_{98\%}\geq$L(%), using different prescription dose levels (PDL) per center (Table 1).

In all centers, a T2-weighted magnetic resonance imaging (MRI) scan and a positron emission tomography (PET) were acquired for each patient to facilitate gross tumour volume (GTV) and OARs delineation. They were rigidly registered to a single-energy computed tomography (CT) scan for planning. Main relevant OARs involved in the clinically used NTCP models i.e., the parotid and submandibular glands, oral cavity and pharyngeal constrictor muscles (PCM), and additional important OARs i.e., spinal cord and the brainstem, were delineated based on international consensus guidelines for CT-based delineation of OARs for HNC patients [21]. For each fraction, daily cone-beam CT (CBCT) are used for pre-treatment position verification and triggering plan adaptations (on repeat CT) if necessary.

Treatment planning

Clinical IMPT treatment plans were generated using the RayStation treatment planning system and transferred to a research system for this study (v11B: center 1, v10B: center 2 and v11A: center 3, RaySearch, Sweden). Planning strategies and beam arrangements were different between the institutions, adapted to each treatment delivery machine and planning technique. Treatment plans from center 1 and center 2 were made using beam angle configurations of 2 anterior (40° and 320°) and 2 posterior (160° and 200°) coplanar oblique beams, with subsequent manual finetuning of the beam angles per patient. Additionally, posterior fields from center 1 plans were divided into range shifted and non-range shifted beams to improve OAR sparing, resulting in 5 to 7 beams per plan. At center 3, plans were generated using beam configurations from 4 up to 11 beams, including field-in-field beams and when needed multiple isocenters if the target exceeded the 20x20 cm field size.

Robustly optimized treatment plans were created using objective-wise worst-case robust optimization modules from RayStation (with setup/range robustness settings SR/RR). 21 scenarios were included for robust optimization, considering geometrical and range scenarios along the three cardinal axes (±SR, 0, 0), (0, ±SR, 0) and (0, 0, ±SR) and the nominal scenario (0, 0, 0) with three relative SPP errors (0, +RR and −RR). At center 1 and center 2, the first chronological 20 patients of the cohort were planned with a SR setting of 5 mm (40/100). Based on retrospective evaluation of treatment accuracy at each of the centers [22], beam quality assurance and pre-treatment position verification carried out at each center, a SR setting of 3 mm was later applied for the rest of patients (40/100). At center 3, a SR setting of 3 mm was used for all plans (20/100). The range robustness (RR) setting was set to 3 % in the 3 centers to handle the relative SPP, in line with [23] and clinically validated in-house in UMCG [24]. Finally, the standardized RE protocol was applied following the DUPROTON consensus guidelines [12]. Adequate CTV and OARs doses were ensured using VWmin and VWmax dose distributions from 28 scenarios based on combinations of geometrical and range shifts according to the SR (±3 mm/±5 mm) settings, and range RR (±3%) settings clinically used, respectively.

Polynomial Chaos Expansion (PCE): PCE-based robustness evaluation

Polynomial Chaos Expansion (PCE) was applied to generate a computationally efficient patient- and treatment plan-specific model of the dose on treatment errors. The dose D_i in a voxel i as affected by a geometrical shift $\vec{\xi} = \left(\xi_x, \xi_y, \xi_z\right)$ and a relative range error ρ was approximated by the series expansion $D_i\left(\vec{\xi}, \rho\right) = \sum_{k=0}^{P} a_{i,k}\Psi_k\left(\vec{\xi}, \rho\right)$, with expansion coefficients $\{a_{i,k}\}$ and multi-dimensional Hermite polynomials $\Psi_k\left(\vec{\xi}, \rho\right)$ [16–20]. For each patient and each treatment plan, a PCE-model of the dose was constructed, allowing the probabilistic robustness evaluation of 100,000 complete fractionated treatments (PCE-based robustness evaluation). Thus, patient- and treatment plan-specific probability distributions of relevant dosimetric parameters (scenario distributions) across the sampled treatments were determined for both CTVs ($CTV_{70.00}$ and $CTV_{54.25}$) and the relevant OARs.

Complete fractionated IMPT treatments were simulated from Gaussian error distributions by drawing a (i) a fixed systematic geometrical (Σ) and a fixed systematic relative range error (ρ) for the complete treatment and (ii) a different random error (σ) for each fraction. Gaussian geometrical systematic and random errors of $\Sigma = 0.92$ mm/1.53 mm (1 SD) and $\sigma = 1.00$ mm/1.67 mm (1 SD) – consistent with a $M=2.5\Sigma + 0.7\sigma = 3$ mm/5 mm based on van Herk's margin recipe and clinical experience – were used for the error scenario sampling [25]. Earlier work [17] showed that different combinations Σ and σ errors that justifies the same SR based on van Herk's margin recipe did not impact the PCE-based robustness evaluations. Range errors were assumed to be purely systematic, with $\rho = 1.5$ % (1 SD) based on literature [26]. A flowchart of the PCE-based robustness evaluation workflow is depicted in Fig. 1.

Evaluation of the CTV dose and TCP between centers

To evaluate the impact of treatment errors, dose to both CTVs was assessed using the probabilistic near minimum and near maximum CTV dose metrics: $D_{99.8\%}$ and $D_{0.2\%}$. These probabilistic metrics were based on prior single-center cross-calibration of the volumes 99.8 % and 0.2 % between protons and photons [18] and subsequently compared with the clinical VWmin-$D_{98\%,CTV}$ and VWmax-$D_{2\%}$ values. Consistency of the CTV dose robustness between centers was evaluated with a comparison of population dose histograms for the $D_{99.8\%}$, $V_{95\%}$ and $D_{0.2\%}$ between the centers. Population dose histograms were determined as the average of the patient-specific probabilities over all patients, derived from the cumulative distributions of the scenario $D_{99.8\%}$, $V_{95\%}$ and $D_{0.2\%}$ distributions (patient probability $D_{99.8\%}$, $V_{95\%}$ and $D_{0.2\%}$ dose histograms), as explained in the supplementary material [SM, section 1]. Inter-patient variation was determined by calculating 5th and 95th percentiles (5th/95th) in each point of the population $D_{99.8\%}$, $V_{95\%}$ and $D_{0.2\%}$ dose histograms. Thus, per center and per SR setting (3 mm/5 mm) population dose histograms for the $D_{99.8\%}$, $V_{95\%}$ and $D_{0.2\%}$ were determined as the patient average across the patient population. For reference, results based on the $D_{98\%}$ and $D_{2\%}$, – commonly used in dose-accumulation studies, but not consistent between photon- and proton-based plan evaluation metrics [18] – are included in the [SM, section 2 and 3]. Per patient and per center evaluation of clinical VWmin-$D_{98\%,CTV}$ with scenario $D_{99.8\%}$ distributions in the CTV and of the probability of achieving $V_{95\%}\geq$99.8 % can be found in [SM, section 4].

Clinical tumour control was based on the TCP modelling developed

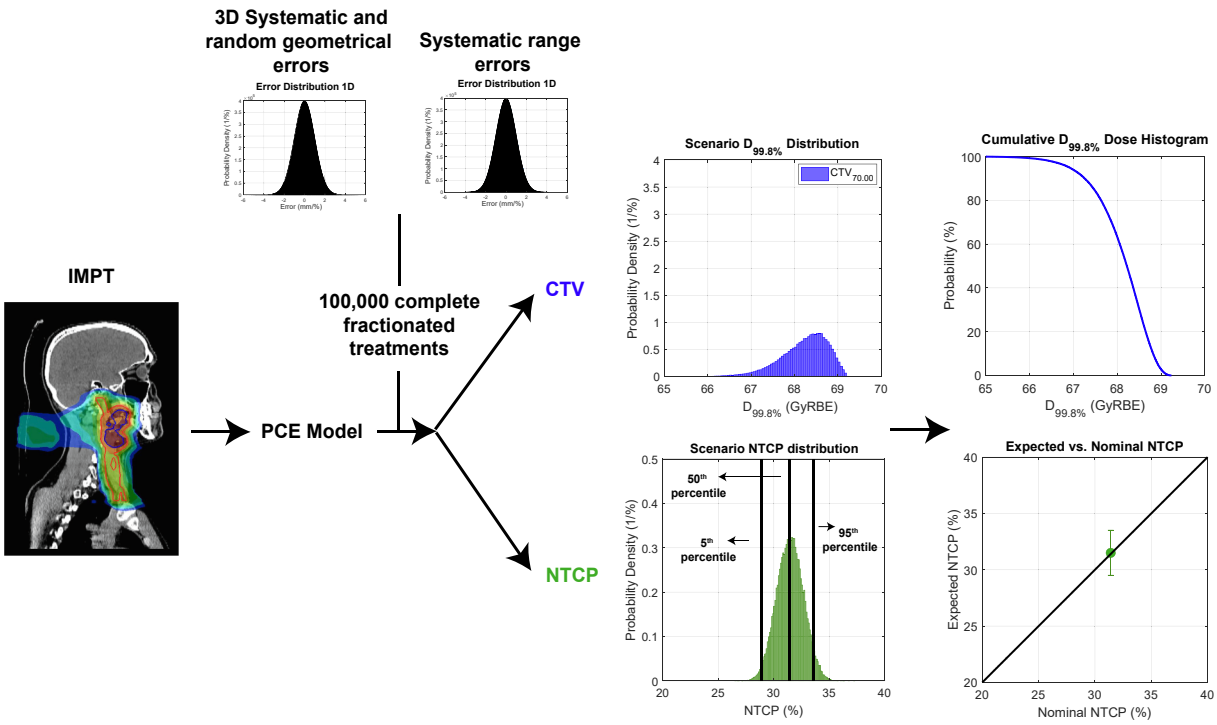

Fig. 1. PCE-based robustness evaluation workflow. First of all, a patient- and treatment plan-specific PCE model of the dose was constructed. Secondly, 100,000 complete fractionated treatments were simulated. Scenario DVH distributions were then calculated according to the simulated treatments. For the CTV, population $D_{99.8\%}$ dose histograms were determined from the averaged cumulative $D_{99.8\%}$ distributions between all patients. For the NTCPs, the 5th/95th percentiles and the expected NTCP were plotted against the nominal NTCP.

by Luhr et al. [27]. The model considers different tumour response for the GTV, $CTV_{70.00}$ and $CTV_{54.25}$, in which the total TCP (TCP_T) is calculated as: $TCP_T = TCP_{GTV} \times TCP_{CTV,70} \times TCP_{CTV,54.25}$, assuming independent TCP probabilities based on the DVH per target region [SM, section 5 and 6]. For further details we refer to the SM from [18]. Two sets of local recurrences probabilities for the different target subvolumes were used for the TCP calculation: (i) 82 %, 16 % and 2 % considered in [28] and (ii) 51.3 %, 29.4 % and 19.3 % considered in a recent study [29], for the GTV, remainder of the $CTV_{70.00}$ and remainder of the $CTV_{54.25}$ respectively. Correlations between expected TCP and nominal TCP were calculated. The robustness of TCP against geometrical and range errors was evaluated via TCP spreads, as the difference between the 95th (down bar) and 5th (upper bar) percentiles of the scenario TCP distribution in [SM, section 6]. The GTV structure of one patient from center 3 was surgically removed before the RT treatment and therefore was excluded from the TCP modelling.

Impact of treatment errors on the MBS protocol

The NTCP modelling from the MBS protocol is based on clinical outcome data from the D_{mean} to OARs, which are sensitive to geometrical and range errors during treatment. Hence, the sensitivity of NTCP was assessed with correlations of the nominal vs. the expected NTCP values calculated from the scenario NTCP distributions on the planning CT. In this study, NTCP models clinically used for two different toxicities were used: risk of grade $\geq$ II and grade $\geq$ III (i) xerostomia and (ii) dysphagia [13–15]. The xerostomia NTCP models are based on the D_{mean} to the parotids and submandibular glands, while the dysphagia NTCP models involved the D_{mean} of the external oral cavity and the PCM. Expected NTCP values were calculated as the average value of the scenario NTCP distributions, calculated from the scenario D_{mean} distributions of the OARs involved in the NTCP models. The robustness of the MBS protocol was evaluated with NTCP spreads (5th/95th percentile range) of the scenario NTCP distribution.

Results

For all 3 centers, population average values of relevant dosimetric parameters ($D_{99.8\%}$,$D_{98\%}$, $D_{2\%}$, $D_{0.2\%}$ and $V_{95\%}$) for the CTV derived with PCE and evaluated at a 90 % population probability are reported and compared with clinical VWmin-$D_{98\%,CTV}$ and VWmax-$D_{2\%,CTV}$ in Table I.

Fig. 2 shows for each center the population $D_{99.8\%}$ dose histograms for both the $CTV_{54.25}$ and $CTV_{70.00}$. For center 1 and 2, the population $D_{99.8\%}$ dose histograms were above 0.95 D_{pres} for both patient cohorts planned with a SR=3 mm and 5 mm. For the SR=3 mm cohort, the population $D_{99.8\%}$ for the $CTV_{70.00}$ and $CTV_{54.25}$ were 67.3 (5th/95th percentiles: 66.9–68.4) GyRBE and 52.2 (51.7–52.6) GyRBE for center 1 (Fig. 2A) and 67.4 (66.5–68.4) GyRBE and 52.3 (51.2–52.8) GyRBE for center 2 (Fig. 2B). For patients planned with a SR=5 mm, values of 67.6 (65.7–68.5) GyRBE and 52.6 (51.9–53.1) GyRBE for center 1 (Fig. 2D) and values of 67.6 (67.3/68.5) GyRBE and 52.3 (50.4/53.2) GyRBE (Fig. 2E) were found respectively. For center 3, the $D_{99.8\%}$ values were partly below the 95 % D_{pres} in that center for both $CTV_{70.00}$ and $CTV_{54.25}$ (Fig. 2C). At a 90 % population probability, the population $D_{99.8\%}$ doses for center 3 were 65.6 (65.1–66.3) GyRBE for the $CTV_{70.00}$, while for the $CTV_{54.25}$, a lower population $D_{99.8\%}$ of 49.3 (47.7–51.4) GyRBE was found for the $CTV_{54.25}$ at a the same population probability, below the planning constraint of 50.5 GyRBE. A more detailed analysis is showed in the [SM, section 4.1 and 4.2]. Population $D_{98\%}$, $D_{0.2\%}$ and $D_{2\%}$ dose histograms and their comparison to the clinical voxel-wise metrics are also displayed in the [SM, section 2 and 3].

These results are confirmed by the population $V_{95\%}$ histograms displayed in Fig. 3. For center 3, a lower mean population $V_{95\%}$ value of 99.52 (98.86–99.89)% and 99.30 (5th/95th percentiles: 98.78–99.75)% were found for the $CTV_{70.00}$ and $CTV_{54.25}$, respectively, lower than the suggested CTV coverage constraint of 99.8 % from the photon plan calibration [18]. A more detailed analysis is presented in the [SM, section 4.3].

Despite this, patients from center 3 showed on average higher

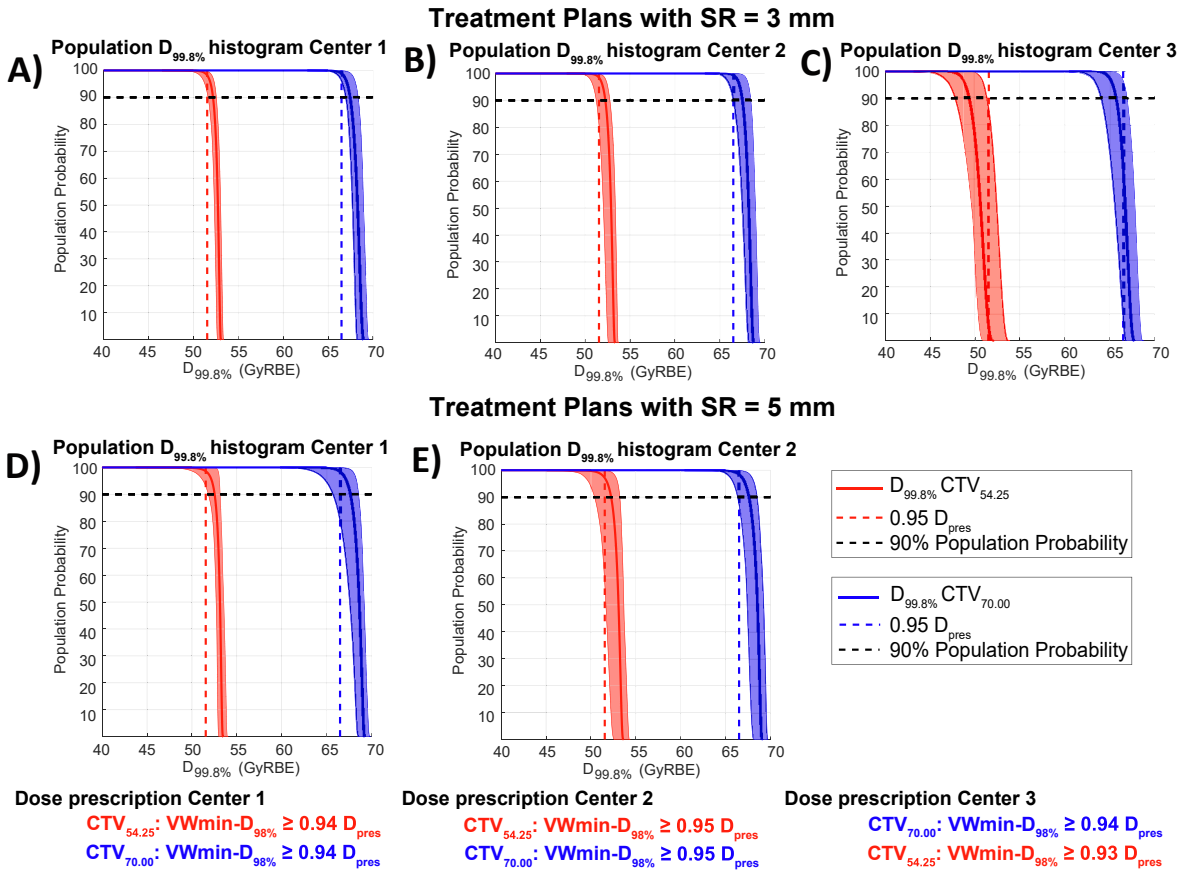

Fig. 2. Population $D_{99.8\%}$ histograms for the $CTV_{70.00}$ (blue) and the $CTV_{54.25}$ (red) for the patients planned with a SR=3 mm (A, B and C) and SR=5 mm (D and E). Inter-patient variations are shown as the 5th/95th percentiles bands respectively, while red and blue solid lines represent the patient average population values. The red and blue dashed lines represent the dose constraints on the $CTV_{54.25}$ and $CTV_{70.00}$ respectively aimed by van Herk's, while the black dashed line corresponds to the 90 % population probability. Prescription doses for the $CTV_{54.25}$ and $CTV_{70.00}$ are reported per center respectively.

population DVH histograms for the GTV, for the remainder of the $CTV_{54.25}$ and for the remainder of the $CTV_{70.00}$ (Fig S9, S10 and S11) and, subsequently, this resulted into higher scenario TCP distributions compared to center 1 and 2 (Fig S12 and S13). The comparison of expected vs. nominal TCP and the evaluation of TCP spreads can be found in [SM, section 5 and 6].

Fig. 4 shows correlation plots of the nominal vs. expected NTCP and their absolute difference (ΔNTCP) for the different toxicities, per center and per SR setting clinically used. A moderate impact on ΔNTCP was found for all 3 centers. For the SR=3 mm patients, ΔNTCP values of 0.03 (range: 0.0/0.1) percentage point (%-point) for center 1, 0.02 (0.0/0.2) %-point for center 2 and 0.02 (0.0/1.4) %-point were found on average for both grade II and III xerostomia and dysphagia complications. Only 1 patient from center 3 showed a maximum ΔNTCP of 0.7 %-point and 1.65 %-point for the xerostomia and dysphagia grade II complications respectively. Slightly larger inter-patient variations in the ΔNTCP were found for the SR=5 mm plans. For these patients, ΔNTCP values of 0.05 (0.0/0.2) %-point and 0.04 (0.0/0.2) %-point were found for center 1 and 2 respectively.

NTCP spreads for the 3 centers showed a moderate impact of geometrical and range errors on NTCPs used for PT patient selection. For the SR=3 mm patient cohort, median NTCP spreads of 3.4 (range:1.2/4.5) %-point and 3.1 (1.0/5.8) %-point for center 1, 2.6 (1.7/3.8) %-point and 3.8 (2.8/5.3) %-point for center 2 and 2.7 (1.1/3.7) %-point and 4.0 (1.6/8.1) %-point for center 3 were found on average for grade II xerostomia and dysphagia complications respectively. For grade III toxicities, median values of 1.0 (0.4/1.3) %-point and 0.7 (0.0/3.0) %-point for center 1, 0.9 (0.6/1.2) %-point and 0.7 (0.5/2.6) %-point for center 2 and 0.9 (0.3/1.2) %-point and 1.1 (0.2/4.3) %-point for center 3 were found respectively. NTCP spreads were larger for the SR=5 mm patients than for the SR=3 mm patients. Only the grade II xerostomia NTCPs of center 1 showed similar differences between both SR settings. For SR=5 mm patients, NTCP spreads of 2.9 (1.0/4.3) %-point and 4.3 (1.7/8.3) %-point for center 1 were found while, for center 2, values of 3.6 (2.5/5.3) %-point and 5.5 (1.4/9.1) %-point resulted respectively for grade II xerostomia and dysphagia toxicity. For grade III toxicities, median values of 1.0 (0.3/1.5) %-point and 1.1 (0.1/3.1) %-point for center 1 and 1.3 (0.9/1.9) %-point and 1.9 (0.4/5.3) %-point were found respectively.

Discussion

In this study, the performance of the Dutch RE and MBS protocols – the first protocols nationally standardized for target dose evaluation and patient selection, respectively, for PT – was evaluated in a multi-institutional setting among the 3 Dutch proton centers. Our main finding is that, after its standardization, their application led to comparable results in terms of TCP and NTCP between the centers, in which a clinically acceptable degree of inter-patient and inter-center variation in target dose and NTCP was observed (Figs. 2 and 4).

For photon treatments, van Herk's margin recipe aimed, with a PTV-$D_{100\%}$ evaluation, for a coverage of 100 % of the CTV with 0.95 D_{pres} with a 90 % population probability. Subsequently, a PTV-$D_{98\%}$ goal was generally adopted by the RT community. Despite this consensus, differences in treatment planning, delivery techniques and local clinical preferences led to inter-institutional and inter-patient variation in target coverage in clinical practice. The aim of the proton RE protocol was to establish an equivalence in terms of CTV dose between PTV-based and

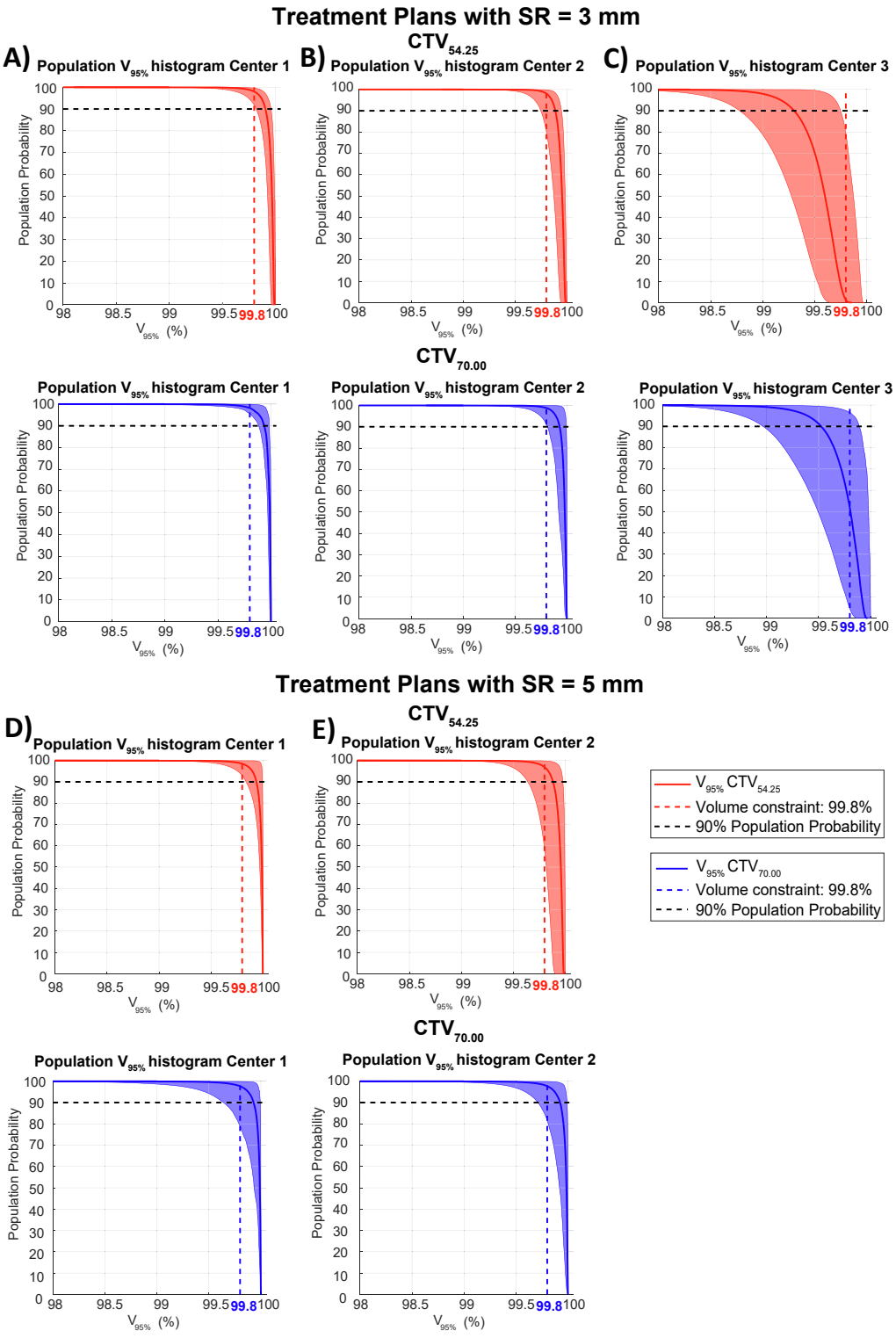

Fig. 3. Population $V_{95\%}$ histograms for the $CTV_{70.00}$ (blue) and the $CTV_{54.25}$ (red) for the patients planned with a SR=3 mm (A, B and C) and SR=5 mm (D and E). Inter-patient variations are shown as the 5th/95th percentiles bands. The red and blue dashed lines represent the probabilistic coverage constraint ($V_{95\%}\geq$99.8 %) on the $CTV_{54.25}$ and $CTV_{70.00}$ respectively, while the black dashed line corresponds to the 90 % population probability.

scenario robust optimization with historical photon treatment planning within an institution, resulting in inter-center differences in the PDLs used in each center.

When correcting for the differences in the clinically accepted VWmin-$D_{98\%,CTV}$, an agreement between the probabilistic CTV dose metric ($D_{99.8\%}$) and the clinical VWmin-$D_{98\%,CTV}$ was found for the $CTV_{70.00}$ in all 3 institutions and for the $CTV_{54.25}$ in center 1 and 2. The inter-patient variation in robust CTV dose can be partially explained by the clinically accepted variation found in the VWmin-$D_{98\%,CTV}$. However, still substantial variation remains, as found for the probabilistic $V_{95\%}$ metric [SM, section 4.3]. For center 3, target coverage is relatively more sensitive to treatment errors than target dose for the $CTV_{70.00}$ and

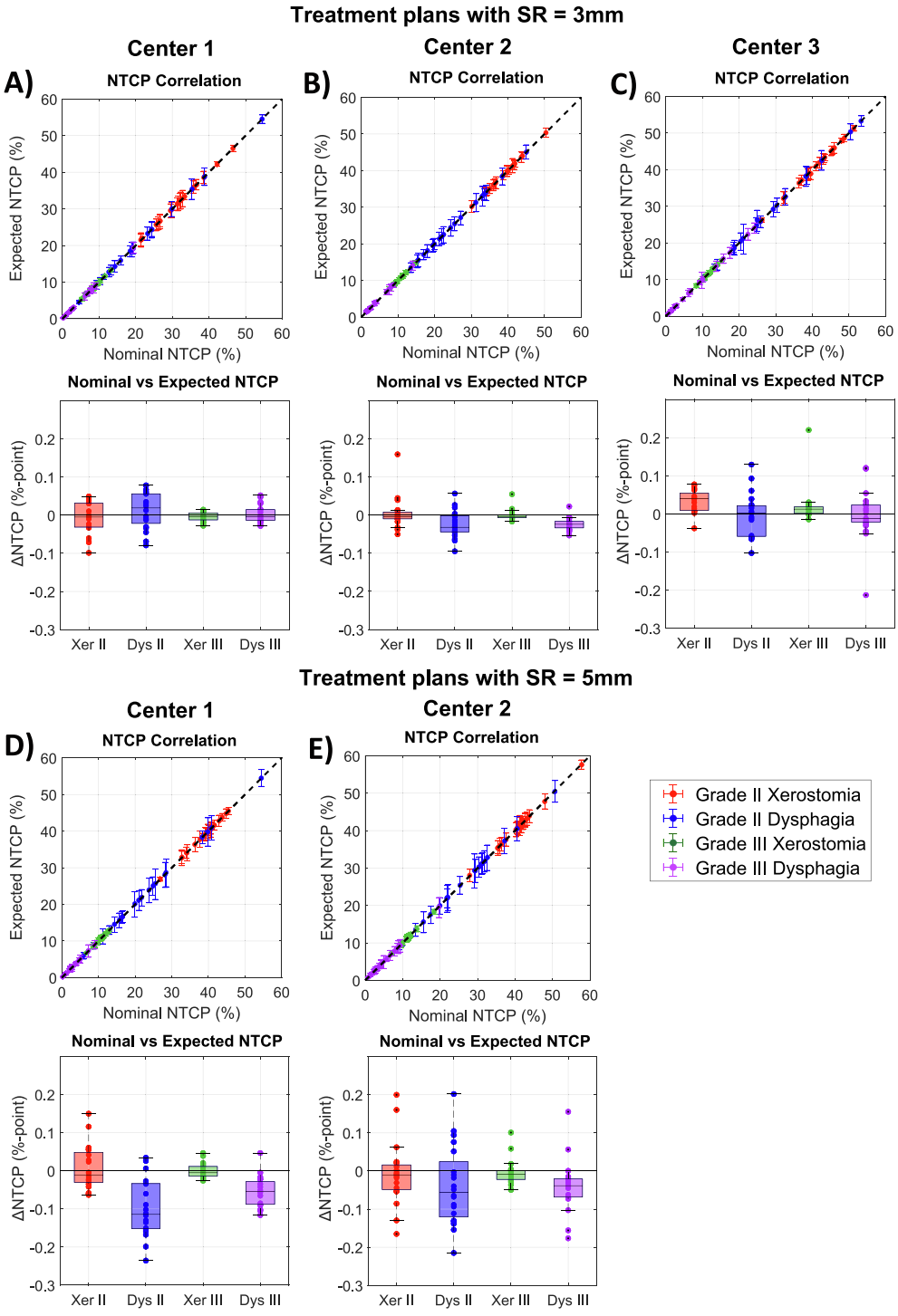

Fig. 4. Expected vs. nominal NTCP correlations for risk of grade II and III xerostomia and dysphagia toxicities, including treatment plans with SR=3 mm (A, B and C) and 5 mm (D, E). Below the correlations, boxplots of the absolute differences between the nominal and expected NTCP (ΔNTCP). The points correspond to the nominal (no setup and range errors) and the expected NTCP values of the sampled treatments, with bars representing the 5th/95th percentiles. The black dashed line corresponds to the identity line.

the $CTV_{54.25}$ compared to center 1 and 2. In particular for the $CTV_{54.25}$ for center 3, a lower agreement between the probabilistic and the clinical metric was found, in which there was a 1 GyRBE reduction on average between the clinical VWmin-$D_{98\%,CTV}$ and the population $D_{99.8\%}$ metric. In this case for the $CTV_{54.25}$, the clinically accepted VWmin-$D_{98\%,CTV}$ values did not ensure probabilistically 0.95 D_{pres} in 99.8 % of the volume [SM, section 4.2]. As depicted in Fig. 3, the $D_{99.3\%}$ could be used as an alternative probabilistic metric to describe the clinical practice at that center.

The lower target coverage in the $CTV_{54.25}$ resulted in a minor impact in terms of modelled TCP, as the systematically higher doses given to the GTV from the higher inhomogeneity dose constraints resulted then into higher TCP_T values compared to center 1 and 2 [SM, section 6].

A limitation of this study may be the lack of consensus of dose metrics for probabilistic robustness evaluation. We chose the probabilistic $D_{99.8\%}$ as a primary dosimetric endpoint based on a prior single-

center calibration against PTV-based photon plans [18], following the premises assumed by van Herk in his margin recipe. However, this probabilistic goal was conservative to describe the clinical practice at center 3 [Fig. 3]. If a relaxed probabilistic $D_{98\%}$ or $D_{99\%}$ metric is used instead of $D_{99.8\%}$, as is common in dose-accumulation studies [30,31], a consistent dose to the CTV is achieved for both the $CTV_{70.00}$ and $CTV_{54.25}$ in all centers [SM, section 2]. For the probabilistic $D_{98\%}$ metric, a lower PDL of 92.7 % D_{pres} in the VWmin dose distribution could be used, in line with [17]. Robustness against treatment uncertainties was similar for both SR=3 and 5 mm cohorts. This reflects that comparable robustness in terms of CTV dose/coverage was maintained after the SR reduction justified by the quality assurance and pretreatment patient verification positioning analysis carried out in each center [18,22]. Additionally, similar inter-patient variation was found between the SR=3 mm and 5 mm plans, underlining that robust optimization works well for both SR robustness settings and that a probabilistic evaluation with PCE could possibly justify margin reduction for these patients if the magnitude of errors is reduced from improved patient-alignment and quality assurance verification protocols, in line with [30].

The RE protocol has 4 limitations. First, it is limited to a fixed small sample of error scenarios. Second, it introduced inter-center variation in PDLs from local PTV-based photon plan calibrations. As a result, these differences have been calibrated into PT clinical practice. Third, similarly to PTV-based evaluations in photon planning, VWmin metrics lack probabilities of CTV dose/coverage for patients in which underdoses of the CTV occurred due to clinical decisions, for instance, to spare OARs. Additionally, for these patients, the robustness of the plan cannot be interpreted, as showed in [32]. Finally, PTV-based and voxel-wise metrics are often used to trigger plan adaptations over the course of the treatment. As found for the $CTV_{54.25}$ in center 3 (Fig S6C), the impact of anatomical variations can compromise clinical VWmin metrics and may not correctly suggest a plan adaptation for a patient that actually qualifies for a plan adaptation, in line with [33].

Probabilistic evaluations could help to solve these limitations. They enable accurate comparisons of robustness across institutions, which could aid in the harmonization and consensus of robustness evaluation protocols, not only at a national, but also at an international level [34]. As a possible next step, dose prescription based on probabilistic CTV dose metrics could reduce the remaining inter-center variations for both photon and proton planning. Additionally, PCE has proven to be an effective and reliable robustness tool as it allowed a more comprehensive understanding of the clinical RE and MBS protocols, enabling to analyze the impact of using different PDLs on CTV dose/coverage and the impact of geometrical and range errors respectively. The combination of PCE-based robustness evaluations with offline and online adaptive radiotherapy strategies could potentially improve current plan adaptation protocols, in which more comprehensive robustness metrics on the CTV need to be determined to assess the suitability of the plans on the patient anatomy of the day.

Another limitation lies in potentially systematic differences between HNC population in the 3 centers, leading to very heterogeneous patient cohorts. As our main goal was to assess the performance of the RE and MBS protocols in different clinical settings, we used clinical IMPT plans of HNC patients treated at these institutions.

We only considered isocentric errors, modelled as rigid shifts and SPP errors. As IMPT dose delivery is local, geometrical shifts, rotations and deformations are approximated to local shifts. Anatomical variations were not included in this analysis, although this analysis could also be performed on verification CTs during the course of the treatment.

For the MBS approach, the small variations found in the ΔNTCP showed that the protocol, which uses the nominal NTCP to qualify patients for PT, is robust and a good proxy of the expected NTCP during treatment. Thus, nominal NTCP could be used to select patients for PT, as it showed ΔNTCP<0.3 %-point and moderate impact of geometrical and range errors. This also demonstrates that the MBS is a flexible and a versatile tool that can be robustly adopted clinically at different PT centers with different planning strategies and different delivery systems. However, NTCP models depend on D_{mean} doses, which are inherently more robust metrics against errors and the impact of anatomical variations was not evaluated. Different results might be expected for NTCP models based on near-maximum dose metrics and if evaluated in repeat CT during the course of a treatment. Similarly, the TCP models and ΔTCP also showed a good correlation between expected and nominal TCP and good robustness against treatment errors (below 0.2 %-point). The NTCP models from the MBS protocol and the TCP models were defined based on correlations between nominal dose to OARs and GTV, $CTV_{54.25}$ and $CTV_{70.00}$, respectively, and clinical toxicity outcomes, which included delivery uncertainties.

Conclusion

After national standardization of the RE and MBS protocol, current clinical practice in the three Dutch proton centers is comparable in modelled clinical endpoints: TCP and NTCP. However, substantial inter-patient and inter-center variation remains in the dosimetric parameters, which can be partially explained by variations in photon and proton clinical practice at each of the centers and by the local calibration of clinical goals recommended in the Dutch protocol for robust target dose evaluation. The MBS approach was moderately impacted by geometrical and range errors and its application in a multi-institutional setting was robust, thus improving the reliability of the Dutch NTCP-based protocols as a referral of patients for PT. Finally, probabilistic robustness evaluations could further lead to: (i) a reduction of inter-center variation in proton and photon planning; (ii) an indication to evaluate a possible margin reduction with organ sparing and (iii) to a better comprehension of treatment plan robustness in clinical practice.

CRediT authorship contribution statement

Jesús Rojo-Santiago: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Steven J.M. Habraken:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Conceptualization. **Mirko Unipan:** Writing – review & editing, Supervision, Resources. **Stefan Both:** Writing – review & editing. **Geert Bosmans:** Writing – review & editing. **Zoltán Perkó:** Writing – review & editing, Supervision, Software, Resources, Methodology. **Erik Korevaar:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Mischa S. Hoogeman:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

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 Kanker Bestrijding (project number 11711).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110441.

References

[1] International Agency for Research on Cancer. Cancer Today. https://gco.iarc.fr/today/home.

[2] Van der Veen J, Nuyts S. Can intensity-modulated-radiotherapy reduce toxicity in head and neck squamous cell carcinoma? Cancers 2017;9:135.

[3] McKeever MR, Sio TT, Gunn GB, Holliday EB, Blanchard P, Kies MS, 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;5:54.

[4] Li X, Lee A, Cohen MA, Sherman EJ, Lee NY. Past, present and future of proton therapy for head and neck cancer. Oral Oncol 2020;110:104879.

[5] Jeans EB, Shiraishi S, Manzar G, Morris LK, Amundson A, McGee LA, et al. An comparison of acute toxicities and patient-reported outcomes between intensity-modulated proton therapy and volumetric-modulated arc therapy after ipsilateral radiation for head and neck cancers. Head Neck 2022;44:359–71.

[6] Ng SP, Pollard III C, Kamal M, Ayoub Z, Garden AS, Bahig H, et al. Risk of second primary malignancies in head and neck cancer patients treated with definitive radiotherapy. npj Precis Oncol 2019;3:22.

[7] Lee JY, Abugharib A, Nguyen R, Eisbruch A. Impact of xerostomia and dysphagia on health-related quality of life for head and neck cancer patients. Expert Review of Quality of Life in Cancer Care 2016;1:361–71.

[8] Water TA, Bijl HP, Schilstra C, Pijls-Johannesma M, Langendijk JA. The potential benefit of radiotherapy with protons in head and neck cancer with respect to normal tissue sparing: a systematic review of literature. Oncologist 2011;16:366–77.

[9] Blanchard, P., Gunn, G. B., Lin, A., Foote, R. L., Lee, N. Y., & Frank, S. J. (2018, January). Proton therapy for head and neck cancers. In *Seminars in radiation oncology* (Vol. 28, No. 1, pp. 53-63). WB Saunders.

[10] Lomax AJ. Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties. Phys Med Biol 2008;53:1027.

[11] 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;53:1043.

[12] Korevaar EW, Habraken SJ, Scandurra D, Kierkels RG, Unipan M, Eenink MG, et al. Practical robustness evaluation in radiotherapy–A photon and proton-proof alternative to PTV-based plan evaluation. Radiother Oncol 2019;141:267–74.

[13] Langendijk JA, Hoebers FJ, De Jong MA, Doornaert P, Terhaard CH, Steenbakkers RJ, et al. National protocol for model-based selection for proton therapy in head and neck cancer. International Journal of Particle Therapy 2021;8:354–65.

[14] Landelijk Platform Protonentherapie, Protonentherapie LP. Landelijk Indicatie Protocol Protonen Therapie Hoofd-halstumoren. 2019.

[15] Tambas M, Steenbakkers RJ, van der Laan HP, Wolters AM, Kierkels RG, Scandurra D, et al. First experience with model-based selection of head and neck cancer patients for proton therapy. Radiother Oncol 2020;151:206–13.

[16] Rojo-Santiago J, Habraken SJ, Lathouwers D, Romero AM, Perkó 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;163:121–7.

[17] Rojo-Santiago J, Habraken SJ, Romero AM, Lathouwers D, Wang Y, Perkó Z, et al. Robustness analysis of CTV and OAR dose in clinical PBS-PT of neuro-oncological tumors: prescription-dose calibration and inter-patient variation with the Dutch proton robustness evaluation protocol. Phys Med Biol 2023;68:175029.

[18] Rojo-Santiago J, Korevaar E, Perkó Z, Both S, Habraken SJ, Hoogeman MS. 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. Radiother Oncol 2023;109729.

[19] Perkó 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.

[20] Perkó Z. Open source generalized Polynomial Chaos Expansion (openGPC) Toolbox. https://gitlab.com/zperko/opengpc.

[21] Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, Grégoire V, 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;117:83–90.

[22] Wagenaar D, Kierkels RG, van der Schaaf A, Meijers A, Scandurra D, Sijtsema M, et al. Robustness evaluation using dose accumulation in head and neck IMPT. Int J Radiat Oncol Biol Phys 2019;105:E744–5.

[23] Meijers A, Free J, Wagenaar D, Deffet S, Knopf AC, Langendijk JA, et al. Validation of the proton range accuracy and optimization of CT calibration curves utilizing range probing. Phys Med Biol 2020;65:03NT02.

[24] Meijers A, Seller Oria C, Free J, Langendijk JA, Knopf AC, Both S. 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;48:1372–80.

[25] Van Herk, M. (2004, January). Errors and margins in radiotherapy. In *Seminars in radiation oncology* (Vol. 14, No. 1, pp. 52-64). WB Saunders.

[26] Taasti VT, Muren LP, Jensen K, Petersen JBB, Thygesen J, Tietze A, et al. Comparison of single and dual energy CT for stopping power determination in proton therapy of head and neck cancer. Physics and imaging in radiation oncology 2018;6:14–9.

[27] Luhr 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.

[28] 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.

[29] Zukauskaite R, Hansen CR, Grau C, Samsøe E, Johansen J, Petersen JB, et al. Local recurrences after curative IMRT for HNSCC: Effect of different GTV to high-dose CTV margins. Radiother Oncol 2018;126:48–55.

[30] Wagenaar D, Kierkels RG, van der Schaaf A, Meijers A, Scandurra D, Sijtsema NM, et al. Head and neck IMPT probabilistic dose accumulation: Feasibility of a 2 mm setup uncertainty setting. Radiother Oncol 2021;154:45–52.

[31] Siang KNW, Both S, Oldehinkel E, Langendijk JA, Wagenaar D. Assessment of residual geometrical errors of clinical target volumes and their impact on dose accumulation for head and neck radiotherapy. Radiother Oncol 2023;188:109856.

[32] Santiago JR, Habraken SJ, Lathouwers D, Romero AM, Perkó Z, Hoogeman MS. PH-0047 CTV and OAR robustness of clinical neuro IMPT: dosimetric impact of the DUPROTON robustness protocol. Radiother Oncol 2021;161:S22–4.

[33] Oud M, Breedveld S, Rojo-Santiago J, Giżyńska MK, Kroesen M, Habraken S, et al. A fast and robust constraint-based online re-optimization approach for automated online adaptive intensity modulated proton therapy in head and neck cancer. Phys Med Biol 2024.

[34] Sterpin E, Widesott L, Poels K, Hoogeman M, Korevaar EW, Lowe M, et al. Robustness evaluation of pencil beam scanning proton therapy treatment planning: A systematic review. Radiother Oncol 2024;110365.