TU Delft

Delft University of Technology

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

Rojo-Santiago, Jesús; Habraken, Steven J.M.; Romero, Alejandra Méndez; Lathouwers, Danny; Wang, Yibing; Perkó, Zoltán; Hoogeman, Mischa S.

DOI
10.1088/1361-6560/acead1

Publication date
2023

Document Version
Final published version

Published in
Physics in medicine and biology

Citation (APA)
Rojo-Santiago, J., Habraken, S. J. M., Romero, A. M., Lathouwers, D., Wang, Y., Perkó, Z., & Hoogeman, M. S. (2023). 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. *Physics in medicine and biology*, *68*(17), Article 175029. https://doi.org/10.1088/1361-6560/acead1

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.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

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

To cite this article: Jesús Rojo-Santiago *et al* 2023 *Phys. Med. Biol.* **68** 175029

View the article online for updates and enhancements.

You may also like

- Deformable image registration to assist clinical decision for radiotherapy treatment adaptation for head and neck cancer patients
Vasiliki Iliadou, Theodore L Economopoulos, Pantelis Karaiskos et al.
- Evaluation of continuous beam rescanning versus pulsed beam in pencil beam scanned proton therapy for lung tumours
Cássia O Ribeiro, Jorvi Terpstra, Guillaume Janssens et al.
- A deep learning-based approach for statistical robustness evaluation in proton therapy treatment planning: a feasibility study
Ivan Vazquez, Mary P Gronberg, Xiaodong Zhang et al.

This content was downloaded from IP address 154.59.124.113 on 05/09/2023 at 15:33

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
23 February 2023

REVISED
12 July 2023

ACCEPTED FOR PUBLICATION
25 July 2023

PUBLISHED
23 August 2023

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

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

Jesús Rojo-Santiago[1,2,*], **Steven J M Habraken**[1,2], **Alejandra Méndez Romero**[1,3], **Danny Lathouwers**[4], **Yibing Wang**[2], **Zoltán Perkó**[4] and **Mischa S Hoogeman**[1,2]

[1] Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, The Netherlands
[2] Department of Medical Physics & Informatics, HollandPTC, Delft, The Netherlands
[3] Department of Radiation Oncology, HollandPTC, Delft, The Netherlands
[4] Delft University of Technology, Department of Radiation Science and Technology, Delft, The Netherlands
[*] Author to whom any correspondence should be addressed.

E-mail: j.rojosantiago@erasmusmc.nl

Keywords: intensity modulated proton therapy, robust treatment planning, robustness evaluation, polynomial chaos expansion, geometrical and range errors, neuro-oncological tumors

Supplementary material for this article is available online

Abstract

Objective. The Dutch proton robustness evaluation protocol prescribes the dose of the clinical target volume (CTV) to the voxel-wise minimum (VWmin) dose of 28 scenarios. This results in a consistent but conservative near-minimum CTV dose ($D_{98\%,CTV}$). In this study, we analyzed (i) the correlation between VWmin/voxel-wise maximum (VWmax) metrics and actually delivered dose to the CTV and organs at risk (OARs) under the impact of treatment errors, and (ii) the performance of the protocol before and after its calibration with adequate prescription-dose levels. *Approach*. Twenty-one neuro-oncological patients were included. Polynomial chaos expansion was applied to perform a probabilistic robustness evaluation using 100,000 complete fractionated treatments per patient. Patient-specific scenario distributions of clinically relevant dosimetric parameters for the CTV and OARs were determined and compared to clinical VWmin and VWmax dose metrics for different scenario subsets used in the robustness evaluation protocol. *Main results*. The inclusion of more geometrical scenarios leads to a significant increase of the conservativism of the protocol in terms of clinical VWmin and VWmax values for the CTV and OARs. The protocol could be calibrated using VWmin dose evaluation levels of 93.0%–92.3%, depending on the scenario subset selected. Despite this calibration of the protocol, robustness recipes for proton therapy showed remaining differences and an increased sensitivity to geometrical random errors compared to photon-based margin recipes. *Significance*. The Dutch proton robustness evaluation protocol, combined with the photon-based margin recipe, could be calibrated with a VWmin evaluation dose level of 92.5%. However, it shows limitations in predicting robustness in dose, especially for the near-maximum dose metrics to OARs. Consistent robustness recipes could improve proton treatment planning to calibrate residual differences from photon-based assumptions.

1. Introduction

Proton therapy (PT) with pencil-beam scanning (PBS) allow us to achieve better dose conformity to the clinical target volume (CTV) compared to conventional radiotherapy (RT) and PT with passive scattering (Bortfeld *et al* 2005, Kosaki *et al* 2012, Langen and Zhu 2018, Florijn *et al* 2020). However, the distribution of pencil-beam Bragg peaks with modulated intensities is very sensitive to errors in beam and patient-alignment (setup or

© 2023 The Author(s). Published on behalf of Institute of Physics and Engineering in Medicine by IOP Publishing Ltd

geometrical error), variations in anatomy and uncertainties in the proton stopping-power prediction (SPP or range error) (Stroom *et al* 1999, van Herk *et al* 2000, Lomax 2008a, 2008b, 2016). These may compromise both organ-at-risk (OAR) sparing and CTV coverage, while conventional expansion margins are not well-suited to mitigate their impact (van Herk *et al* 2004, Unkelbach *et al* 2018). To this end, scenario-based robust minimax optimization (Fredriksson *et al* 2011, Unkelbach *et al* 2007, 2018) and the robustness evaluation (Henríquez and Castrillón 2008, Korevaar *et al* 2019, Buti *et al* 2020, Hernandez *et al* 2020, Teoh *et al* 2020, Sterpin *et al* 2021, Rojo-Santiago *et al* 2021a) are widely used in PBS-PT nowadays. Both the optimization and evaluation are based on a sample set of (combined) geometrical and SPP (range) error scenarios, replacing planning target volume (PTV) margins (Liu *et al* 2013a, 2013b, van Dijk *et al* 2016).

In the Netherlands, a national proton robustness evaluation protocol has been established following the Dutch Proton Therapy (DUPROTON) group guidelines (Korevaar *et al* 2019). A voxel-wise minimum (VWmin) dose level is prescribed to the CTV, while near-maximum doses to the CTV and serial OARs are assessed on a voxel-wise maximum (VWmax) dose distribution. This protocol was defined in order to establish a robustness evaluation for PT consistent with PTV-based photon plan evaluation metrics. Although it has been in use in all three Dutch proton therapy centers since 2018, it has some known limitations. In a recent paper (Rojo-Santiago *et al* 2021a), a probabilistic robustness evaluation using polynomial chaos expansion (PCE) for a cohort of neuro-oncological patients was performed. It was found that the DUPROTON robustness evaluation protocol, which uses 28 error scenarios, is safe but conservative in terms of dose delivered to the CTV. These results indicate the following:

(i) A dosimetric calibration of the robustness evaluation protocol is required. The conservativism of the protocol can partially be explained by the construction of the VWmin dose as a composite of extreme scenario voxel doses. The fact that setup robustness settings are often derived from photon-based margin recipes, while the underlying assumption of the static dose cloud approximation does not hold for PT, may also play a role.

(ii) The consistency of the protocol needs to be assessed. Neither the degree of inter-patient variation in the protocol, nor how that depends on the number and (sub)set of scenarios used for the evaluation is known.

(iii) It is unknown how clinical VWmax dose metrics correlate with delivered dose to serial OARs. In the DUPROTON consensus paper (Korevaar *et al* 2019), it was found that the clinically used VWmax-$D_{2\%,CTV}$ to the CTV is conservative by 2.3 percentage points (p.p.), but no data are available for serial OARs.

To address the abovementioned points, we systematically and quantitatively investigated the robustness of dose to the CTV and serial OARs, for a cohort of 21 clinically robust neuro-oncological treatment plans. The impact of geometrical and range errors was modeled with PCE. PCE was applied to perform a robustness evaluation with 100,000 complete fractionated treatments per plan and naturally results in proper statistical weighting of the scenarios. Treatment courses were sampled from error distributions consistent with van Herk's photon-based margin recipe (van Herk *et al* 2000), also used as the basis of the Dutch proton robustness evaluation protocol. First, we analyzed (i) how the clinically used near-minimum ($D_{98\%}$) VWmin and near-maximum ($D_{2\%}$ and $D_{0.03cc}$) VWmax metrics correlate with corresponding evaluation metrics in delivered dose to the CTV and serial OARs. Second, (ii) how the Dutch proton robustness evaluation protocol can be calibrated in terms of dose for different scenario subsets (Korevaar *et al* 2019) and what degree of inter-patient variation remains. Finally, (iii) how a probabilistically derived robustness recipe, consistent with the requirements van Herk used in his derivation (van Herk *et al* 2000) and derived from this clinical cohort with the clinical treatment planning software (TPS), differs from the photon-based margin recipe before and after calibration of the protocol.

2. Method

2.1. Patient data and treatment planning

The first 21 neuro-oncological patients treated at our center for meningioma, grade-I glioma, grade II–III oligodendroglioma with 1p/19q co-deletion and grade-II astrocytoma with isocitrate dehydrogenase (IDH) mutation and robustly planned according to clinical protocol, were analyzed (van der Weide *et al* 2021). The prescribed doses (D_{pres}) were 45 Gy(RBE) (1 case), 50.4 Gy(RBE) (15 cases), 54 Gy(RBE) (2 cases) and 59.4 Gy(RBE) (3 cases) in 1.8 Gy(RBE) fractions, prescribed to the VWmin dose of 28 evaluation scenarios (see figure 1). Planning goals for the CTV were specified on the VWmin near-minimum dose (VWmin-$D_{98\%,CTV} \geqslant$ 95% D_{pres}) and on the VWmax near-maximum CTV dose (VWmax-$D_{2\%,CTV} \leqslant$ 107% D_{pres}) (ICRU 1993, 1999). Furthermore, planning constraints in the VWmax-$D_{0.03cc,OARs}$ and on the VWmax-$D_{mean,OARs}$ for the

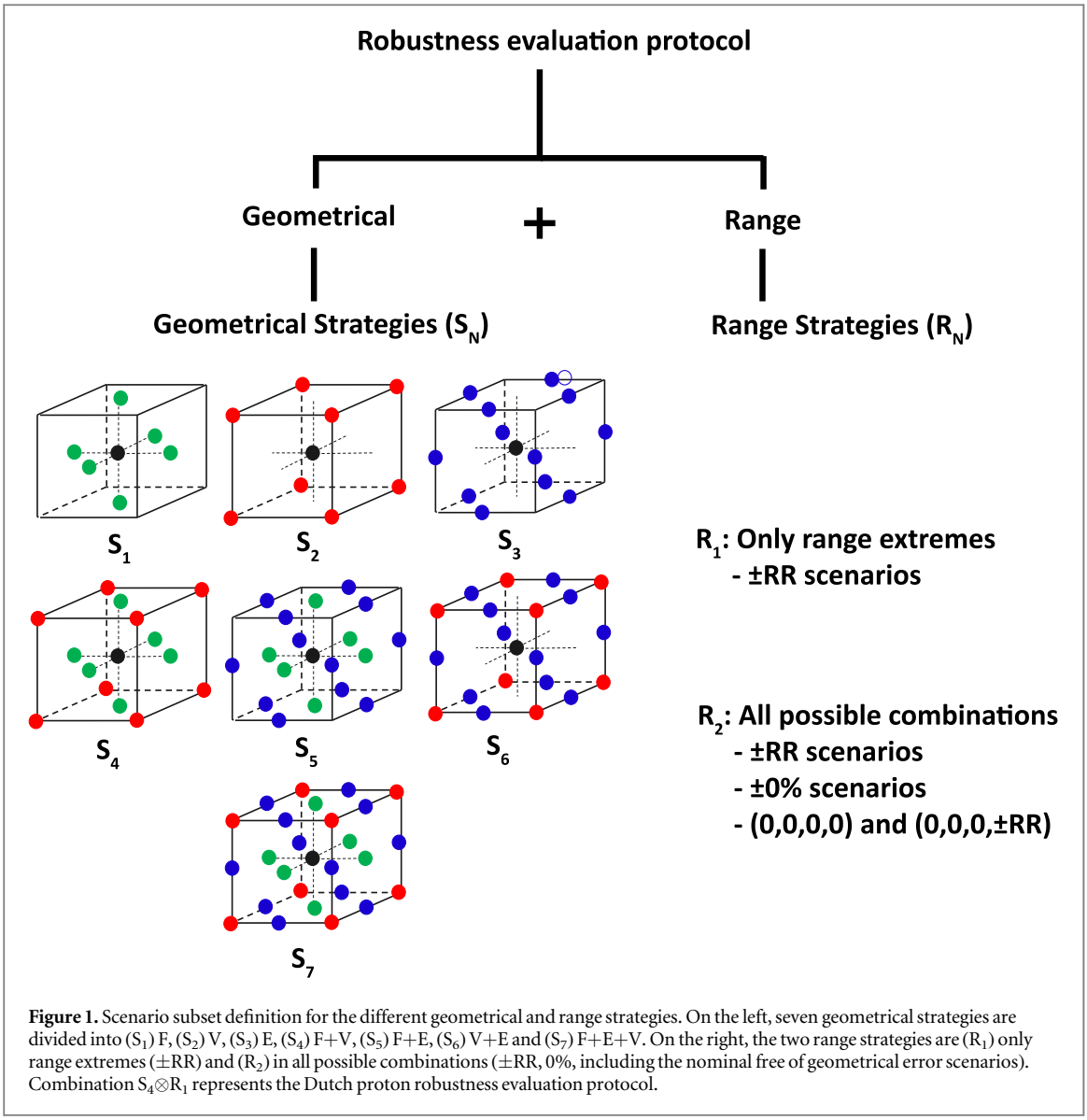

Figure 1. Scenario subset definition for the different geometrical and range strategies. On the left, seven geometrical strategies are divided into (S_1) F, (S_2) V, (S_3) E, (S_4) F+V, (S_5) F+E, (S_6) V+E and (S_7) F+E+V. On the right, the two range strategies are (R_1) only range extremes (±RR) and (R_2) in all possible combinations (±RR, 0%, including the nominal free of geometrical error scenarios). Combination $S_4 \otimes R_1$ represents the Dutch proton robustness evaluation protocol.

relevant serial OARs (Eekers *et al* 2018, Weide *et al* 2020) were also included. A constant relative biological effectiveness (RBE) of 1.1 was assumed. For more details, we refer to (Rojo-Santiago *et al* 2021a). All treatment plans were made using RayStation (version 7, RaySearch Labs, Sweden) TPS, with patient-specific non-coplanar arrangements of two or three beam directions. They were made using minimax robust optimization (Fredriksson *et al* 2011, Unkelbach *et al* 2007, 2018) and evaluated with VWmin and VWmax dose distributions of 28 evaluation scenarios (Korevaar *et al* 2019). An isotropic setup robustness (SR) setting of 3 mm was used to account for geometrical errors. Based on errors in the conversion of the CT number to proton stopping-power ratio from the literature (Lomax 2008a, van der Voort *et al* 2016), a relative range robustness (RR) setting of 3% was used, i.e. uncertainties of ±3% were taken into account.

2.2. Scenario subsets with the DUPROTON protocol

With the Dutch proton robustness evaluation protocol (DUPROTON protocol), voxel-wise dose distributions from 28 evaluation scenarios are generated to assess clinical planning goals. As maximum and minimum voxel dose levels of all scenarios are considered in this approach, different scenario selections will result in different VWmin/max dose distributions. To analyze this dependence, different sets of geometrical (seven geometrical strategies S_N, see figure 1) and range error scenarios (two range strategies R_N), all within the framework of the DUPROTON protocol were combined (scenario subsets $S_N \otimes R_N$) (Korevaar *et al* 2019). As depicted in figure 1, 14 subsets of a total of 81 error scenarios were defined. The S_N geometrical error scenarios were selected as the normalized vectors, to the clinical SR setting used, pointing towards the faces (Fs), vertices (Vs) and edges (Es) of a cube. This resulted in seven geometrical strategies, which were ordered according to the number of error scenarios included: (S_1) F (six error scenarios); (S_2) V (eight error scenarios); (S_3) E (12 error scenarios); (S_4) F+V

(14 error scenarios); (S_5) F+E (18 error scenarios); (S_6) V+E (20 error scenarios); (S_7) F+V+E (26 error scenarios). The R_N range error scenarios were selected following two different strategies: (R_1) ±3% range extremes and (R_2) ±3%, 0% and also the nominal (free of geometrical error) scenarios for each of the RR values. The scenario subset that is calibrated in the DUPROTON protocol and clinically used in the three Dutch proton centers, results from the combination of the geometrical strategy S_4 and range strategy R_1 ($S_4 \otimes R_1$).

2.3. PCE-based robustness evaluation

PCE was applied to provide a computationally efficient patient- and treatment plan-specific analytical model of the dependence of voxel doses on treatment uncertainties. In a 3D dose distribution, the dose D_i of each voxel i is approximated by 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 $\Psi_k(\vec{\xi},\ \rho)$, expressing the dose affected by a geometrical shift $\vec{\xi} = (\xi_x,\ \xi_y,\ \xi_z)$ and a relative range error ρ (Le Maître and Knio 2010, Perkó *et al* 2016). The expansion coefficients $\{a_{i,k}\}$ are approximated by regression and the number of polynomial terms and regression points are selected to find the optimum between model accuracy and computational time. PCE enables the sampling of 100,000 complete fractionated treatments with proper statistical weighting, calculating the corresponding dose distributions in approximately a millisecond per scenario (PCE-based robustness evaluation). For the validation on the current application in neuro-oncological targets and more technical details, we refer to (Rojo-Santiago *et al* 2021a).

For the PCE-based robustness evaluation, treatment courses were sampled assuming systematic and random geometrical (Σ and σ) and systematic range (ρ) errors (1 SD) from Gaussian distributions. The (1 SD) errors were chosen since they exactly match a 3 mm SR in treatment planning, given by the linearized photon-based margin recipe $M = 2.5\Sigma + 0.7\sigma$, and to be consistent with clinical experience. Thus, a systematic and a random geometrical error of $\Sigma = 0.92$ mm and $\sigma = 1.00$ mm were considered for the PCE-based robustness evaluation. For more error combinations, see [Supplementary Material (SM), section S1]. For the range error, a fixed systematic SPP value of 1.2% ± 1.0% (1 SD) was used for the PCE-based robustness evaluations (Wohlfahrt *et al* 2017, 2018, 2019). Thus, one systematic geometrical and one systematic range error were sampled for each treatment course and one random geometrical error for each treatment fraction. Using PCE, scenario probability distribution of voxel doses and clinically relevant dose-volume histogram (DVH) parameters for the CTV (PCE-$D_{98\%,CTV}$ and PCE-$D_{2\%,CTV}$) and for the serial OARs (PCE-$D_{0.03cc,OARs}$) were obtained per patient.

2.4. Calibration of the DUPROTON protocol

For all 14 $S_N \otimes R_N$ scenario subsets, VWmin-$D_{98\%,CTV}$, VWmax-$D_{2\%,CTV}$ and VWmax-$D_{0.03cc,OARs}$ values were derived for the CTV and for the main serial OARs (brainstem and the optic system), respectively. In order to see how these metrics translate into delivered dose, scenario $D_{98\%,CTV}$, $D_{2\%,CTV}$ and $D_{0.03cc,OARs}$ distributions (PCE-$D_{98\%,CTV}$, PCE-$D_{2\%,CTV}$ and PCE-$D_{0.03cc,OARs}$) were compared against their corresponding VWmin-$D_{98\%,CTV}$, VWmax-$D_{2\%,CTV}$ and VWmax-$D_{0.03cc,OARs}$ values, for all scenario subsets. For the CTV, the 50th (median) and 90th percentiles of the scenario $D_{98\%,CTV}/D_{2\%,CTV}$ were linearly correlated against the VWmin-$D_{98\%,CTV}/D_{2\%,CTV}$ doses. For the serial OARs, non-linear regression models ($y = Ax^2+Bx$) of the 50th (median) and the 98th percentiles of the scenario $D_{0.03cc,OARs}$ distributions against the clinical VWmax-$D_{0.03cc,OARs}$ were derived.

To calibrate the protocol in terms of CTV dose, VWmin-$D_{98\%,CTV}$ and VWmax-$D_{2\%,CTV}$ doses were scaled to a fixed percentile of the scenario $D_{98\%,CTV}$ distribution. In line with van Herk (van Herk *et al* 2000), they were consistently scaled per patient to achieve at least 95% of D_{pres} at the 90th percentile of the $D_{98\%,CTV}$ probability distribution (10th percentile PCE-$D_{98\%,CTV} = 95\%D_{pres}$). Furthermore, scaled VWmin-$D_{98\%,CTV}$ and VWmax-$D_{2\%,CTV}$ boxplots were generated for all 14 scenario subsets. Adequate prescription-dose levels (L) for all scenario subsets within the protocol were determined by evaluating the median of the scaled VWmin-$D_{98\%,CTV}$ values (L ($S_N \otimes R_N$)).

2.5. Comparison of the robustness recipe with the photon-based margin recipe

Since the assumptions underlying the static dose cloud approximation do not apply to PBS-PT, photon-based margin recipes cannot be directly applied to calculate the SR setting. To this end, PCE was used to construct a robustness recipe, which amounts to the different combinations of systematic (Σ) and random (σ) geometrical errors for which adequate CTV dose with a pre-defined probability is exactly achieved with the clinical 3 mm SR setting. The probability of achieving adequate CTV dose was defined as the probability of meeting the planning CTV constraint ($D_{98\%,CTV} \geqslant 0.95\ D_{pres}$) for a given percentile of the scenario $D_{98\%,CTV}$ distribution. Therefore, robustness recipes aiming to achieve adequate CTV dose for the 10th (90% robustness recipe), 5th (95% robustness recipe) and 2nd (98% robustness recipe) percentiles of the $D_{98\%,CTV}$ scenario (and population) distribution were derived. For an initial combination of Σ and σ geometrical errors, PCE was first used to sample 100,000 fractionated treatments to determine the $D_{98\%,CTV}$ scenario distribution for all 21 plans. For each of the

21 $D_{98\%,CTV}$ distributions, the probability of achieving adequate dose was determined as the probability of meeting the planning CTV constraint $P_{const} = P(D_{98\%,CTV} \geqslant 0.95\, D_{pres})$. If the averaged probability for all 21 plans did not meet the criterion with a bandwidth of 0.1 p.p., the value of the geometrical Σ was iteratively changed. A non-linear three parameter function was used to fit the recipes: $\Sigma = -a\sigma/\exp(-b\sigma^2) + c$. The coefficients for each of the recipes are tabulated in [SM, section S2].

Robustness recipes for two different situations were determined. The first situation addresses the remaining differences between photon- and proton-based robustness recipes after calibration of the DUPROTON protocol (robustness recipe after protocol calibration). To this end, treatment plans for all patients were scaled according to the VWmin adequate dose evaluation level (L) of the scenario subset used clinically in the DUPROTON protocol ($L(S_4 \otimes R_1)$), determined in section 2.4. The second situation focuses on how the protocol performs when SR settings are tight against the errors assumed. Thus, no protocol calibration was used for the derivation of this recipe (without protocol calibration) and the treatment plans were scaled per patient to achieve the D_{pres} in the 50th percentile of the scenario $D_{50\%,CTV}$ distribution (50th percentile PCE-$D_{50\%,CTV} = D_{pres}$) to reduce inter-patient variation.

To assess the applicability of the robustness recipe, different combinations of geometrical Σ and σ errors satisfying the $P_{const} = 90\%$ and 98% robustness recipe were evaluated and compared against the photon-based margin recipe in the [SM, section S1]. The evaluation of extreme zones of the clinical and photon-based recipes (where Σ or σ are 0 mm) were excluded since they are not realistic in clinical practice.

2.6. Statistical analysis

A statistical analysis on the median (Wilcoxon signed-rank test) and on the data dispersion (Ansari–Bradley test) were performed using Matlab (Mathworks version R2017a) to evaluate the differences between the scenario subsets. A p-value < 0.05 was considered to be statistically significant.

3. Results

3.1. Correlation between clinical plan evaluation metrics versus probabilistic CTV dose metrics

In order to assess differences of the protocol in the selection of the scenario subset, VWmin/VWmax CTV and VWmax OARs dose values for the different combinations of geometrical (S_1 to S_7) and range (R_1 or R_2) scenarios subsets were compared to actual delivered CTV ($D_{98\%,CTV}$, $D_{2\%,CTV}$) and OARs ($D_{0.03cc,OARs}$) dose metrics. The coefficients of the linear and non-linear regressions of PCE against the clinical CTV and OAR voxel-wise metrics can be found in table 1 and table 2, respectively. The inclusion of more geometrical scenarios in the subsets leads to a decrease in VWmin-$D_{98\%,CTV}$ and an increase in VWmax-$D_{2\%,CTV}$ values ($p < 0.05$), increasing the conservativism of the protocol. At a 10th percentile of the scenario $D_{98\%,CTV}$ distribution, a significant increase in the slope from 1.022 (S_1) to 1.030 (S_7) was found, while, for R_2, a value from 1.023–1.031 was obtained ($p < 0.05$). For the 90th percentile of the $D_{2\%,CTV}$ distribution, an increase of 0.6 percentage points (p.p.) and 0.5 p.p. along the geometrical strategies were found for R_1 and R_2, respectively. For the OARs, the correlation was the best for the zero-dose and the high-dose region, with the largest PCE-$D_{0.03cc,OARs}$ and VWmax-$D_{0.03cc,OARs}$ at 60% of the D_{pres} (figure 2(b)). Non-linear coefficients A and B were on average 0.35 and 0.61 for the 98th percentile fitting ($R^2 = 0.99$), which respectively increased and decreased with the addition of geometrical scenarios. Despite the considerable difference in the number of scenarios between R_1 and R_2 strategies, statistically non-significant ($p > 0.05$) differences in the slopes and non-linear coefficients were found between range strategies. Correlation of the voxel-wise CTV and OARs dose metrics for the scenario subset clinically used in the DUPROTON protocol ($S_1 \otimes R_4$) are depicted in figure 2. A visualization of the conservativism of the DUPROTON protocol on these metrics can be found in figure 3.

3.2. Adequate prescription-dose evaluation levels for the DUPROTON protocol

Differences between VWmin-$D_{98\%,CTV}/D_{pres}$ and VWmax-$D_{2\%,CTV}/D_{pres}$ depending on the geometrical (S_1 to S_7) and range strategies (R_1 and R_2) are displayed in figure 4. Dose metrics were scaled for each patient to the 10th percentile of their scenario $D_{98\%,CTV}$ distribution to determine adequate dose evaluation levels for each scenario subset. All scaled VWmin-$D_{98\%,CTV}/D_{pres}$ extended below the target clinical criteria ($D_{98\%,CTV} \geqslant 95\%D_{pres}$). Assuming a population coverage probability of 90%, adequate dose evaluation levels from 93.0% (figure 4(a): $S_1 \otimes R_1$) to 92.2% (figure 4(b): $S_7 \otimes R_2$) on average were found compared to the clinically used 95%. The protocol also results in more homogeneous plans than expected, in which scaled VWmax-$D_{2\%,CTV}/D_{pres}$ values of 1.01 ($S_1 \otimes R_1$) to 1.02 ($S_7 \otimes R_2$) on average were found. Inter-patient variation had a larger impact on the clinical VWmin-$D_{98\%,CTV}/D_{pres}$ than inter-scenario subset variation, where no significant differences resulted for the latter ($p > 0.05$). Further analysis based on other combinations of σ and Σ errors can be found in the [SM, section S1].

Table 1. Linear regression coefficients for the percentiles of the scenario $D_{98\%,CTV}$ (50th and 10th percentiles) and $D_{2\%,CTV}$ (50th and 90th percentiles) distributions against the clinical CTV voxel-wise values. The table shows the slopes for all scenario subsets with the 95% confidence interval in brackets.

CTV	PCE percentiles versus voxel-wise correlation			
	50th PCE-$D_{98\%,CTV}$-VWmin-$D_{98\%,CTV}$		50th PCE-$D_{2\%,CTV}$-VWmax-$D_{2\%,CTV}$	
S_N	R_1	R_2	R_1	R_2
S_1	1.026 (1.024–1.028)	1.027 (1.025–1.029)	0.982 (0.981–0.984)	0.981 (0.980–0.982)
S_2	1.028 (1.026–1.029)	1.029 (1.027–1.031)	0.981 (0.980–0.983)	0.979 (0.978–0.981)
S_3	1.030 (1.028–1.032)	1.031 (1.029–1.033)	0.980 (0.978–0.981)	0.978 (0.977–0.980)
S_4	1.031 (1.029–1.033)	1.032 (1.030–1.034)	0.979 (0.978–0.980)	0.978 (0.976–0.979)
S_5	1.032 (1.030–1.034)	1.033 (1.031–1.035)	0.978 (0.977–0.979)	0.977 (0.976–0.978)
S_6	1.032 (1.030–1.034)	1.033 (1.031–1.035)	0.978 (0.976–0.979)	0.977 (0.975–0.978)
S_7	1.034 (1.031–1.036)	1.035 (1.032–1.037)	0.977 (0.975–0.979)	0.976 (0.974–0.977)
Average	1.030 (1.028–1.034)	1.031 (1.029–1.033)	0.979 (0.978–0.981)	0.978 (0.977–0.979)
	PCE percentiles versus voxel-wise correlation			
	10th PCE-$D_{98\%,CTV}$-VWmin-$D_{98\%,CTV}$		90th PCE-$D_{2\%,CTV}$-VWmax-$D_{2\%,CTV}$	
S_N	R_1	R_2	R_1	R_2
S_1	1.022 (1.020–1.023)	1.023 (1.022–1.025)	0.985 (0.983–0.986)	0.983 (0.982–0.984)
S_2	1.023 (1.022–1.025)	1.025 (1.023–1.026)	0.984 (0.982–0.985)	0.982 (0.981–0.983)
S_3	1.026 (1.024–1.028)	1.027 (1.025–1.029)	0.982 (0.981–0.983)	0.981 (0.979–0.982)
S_4	1.027 (1.025–1.028)	1.028 (1.026–1.029)	0.981 (0.980–0.983)	0.980 (0.979–0.982)
S_5	1.028 (1.026–1.030)	1.029 (1.027–1.031)	0.980 (0.979–0.982)	0.979 (0.978–0.981)
S_6	1.028 (1.026–1.030)	1.029 (1.027–1.031)	0.980 (0.979–0.982)	0.979 (0.977–0.981)
S_7	1.030 (1.028–1.032)	1.031 (1.029–1.033)	0.979 (0.976–0.981)	0.978 (0.976–0.980)
Average	1.026 (1.024–1.028)	1.027 (1.036–1.029)	0.982 (0.980–0.983)	0.980 (0.979–0.982)

Table 2. Second-order regression coefficients for the percentiles of the scenario $D_{0.03cc,OARs}$ (50th and 98th percentiles) distributions against the clinical OARs voxel-wise values. The table shows the values of the non-linear coefficients A and B for all scenario subsets with the 95% confidence interval in brackets.

OARs	50th percentile PCE-$D_{0.03cc}$-VWmax-$D_{0.03cc}$			
	R_1		R_2	
S_N	A	B	A	B
S_1	0.69 (0.63–0.75)	0.25 (0.19–0.31)	0.68 (0.62–0.74)	0.25 (0.20–0.31)
S_2	0.73 (0.68–0.79)	0.20 (0.15–0.25)	0.73 (0.68–0.78)	0.20 (0.15–0.25)
S_3	0.75 (0.69–0.80)	0.18 (0.13–0.23)	0.75 (0.69–0.80)	0.18 (0.13–0.23)
S_4	0.74 (0.68–0.80)	0.19 (0.13–0.24)	0.74 (0.68–0.79)	0.19 (0.13–0.24)
S_5	0.75 (0.70–0.81)	0.17 (0.12–0.23)	0.75 (0.69–0.81)	0.17 (0.12–0.23)
S_6	0.75 (0.70–0.81)	0.17 (0.12–0.22)	0.75 (0.69–0.80)	0.18 (0.12–0.23)
S_7	0.76 (0.70–0.81)	0.17 (0.11–0.22)	0.75 (0.69–0.81)	0.17 (0.11–0.22)
Average	0.74 (0.68–0.80)	0.19 (0.14–0.24)	0.74 (0.68–0.79)	0.19 (0.14–0.24)
	98th percentile PCE-$D_{0.03cc}$-VWmax-$D_{0.03cc}$			
	R_1		R_2	
S_N	A	B	A	B
S_1	0.29 (0.24–0.34)	0.68 (0.63–0.72)	0.29 (0.24–0.33)	0.68 (0.64–0.73)
S_2	0.34 (0.30–0.38)	0.62 (0.58–0.66)	0.34 (0.30–0.38)	0.62 (0.59–0.66)
S_3	0.36 (0.33–0.40)	0.60 (0.56–0.64)	0.36 (0.32–0.40)	0.60 (0.56–0.64)
S_4	0.36 (0.31–0.40)	0.61 (0.56–0.65)	0.35 (0.31–0.40)	0.61 (0.56–0.65)
S_5	0.37 (0.33–0.41)	0.59 (0.55–0.63)	0.37 (0.33–0.41)	0.59 (0.55–0.63)
S_6	0.37 (0.34–0.41)	0.59 (0.55–0.62)	0.37 (0.33–0.40)	0.59 (0.55–0.63)
S_7	0.38 (0.34–0.42)	0.58 (0.54–0.62)	0.37 (0.33–0.41)	0.58 (0.54–0.62)
Average	0.35 (0.31–0.39)	0.61 (0.57–0.65)	0.35 (0.31–0.39)	0.61 (0.57–0.65)

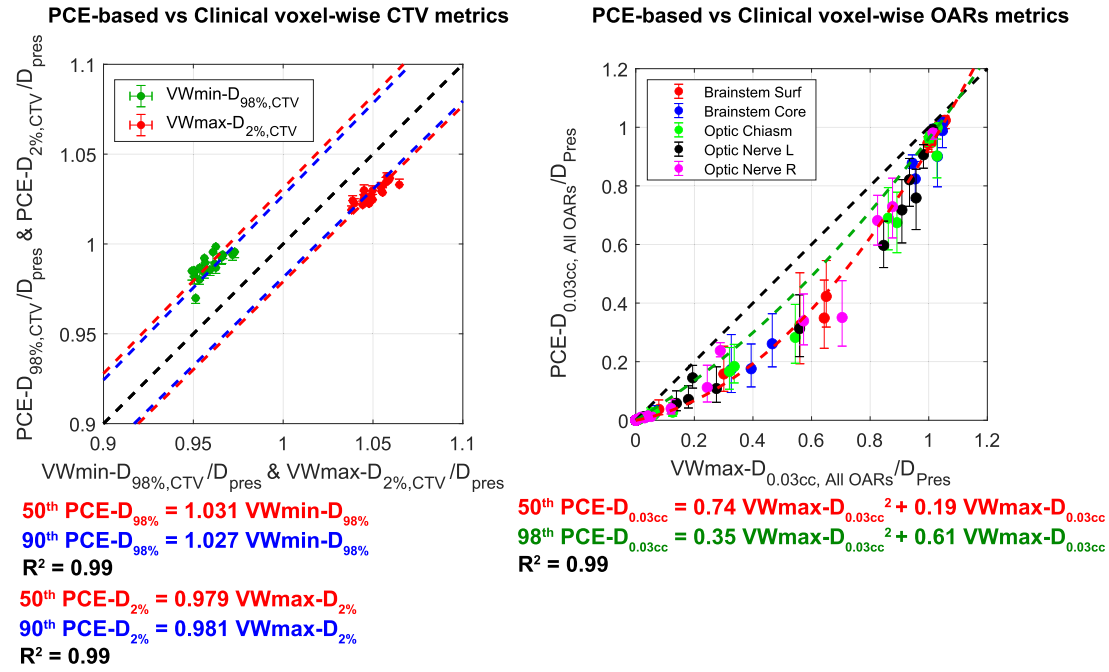

Figure 2. Comparison of clinical voxel-wise metrics with PCE scenario distributions for the clinical robustness evaluation protocol ($S_4 \otimes R_1$). On the left, correlations for the CTV dose metrics. Points correspond to the median PCE-$D_{98\%,CTV}$ and PCE-$D_{2\%,CTV}$ values, and the 90th and 10th percentiles are represented by the up and down error bars, respectively. Identity line (black), the 90% (blue) and the 50% (red) linear regressions are also displayed ($R^2 = 0.99$). On the right, correlations for the OARs dose metrics. Points correspond to the median PCE-$D_{0.03cc,OARs}$ value, while the 98th and 2th percentiles are represented by the up and down error bars. The 98% (dark green) and the 50% (red) non-linear regressions are also displayed ($R^2 = 0.99$).

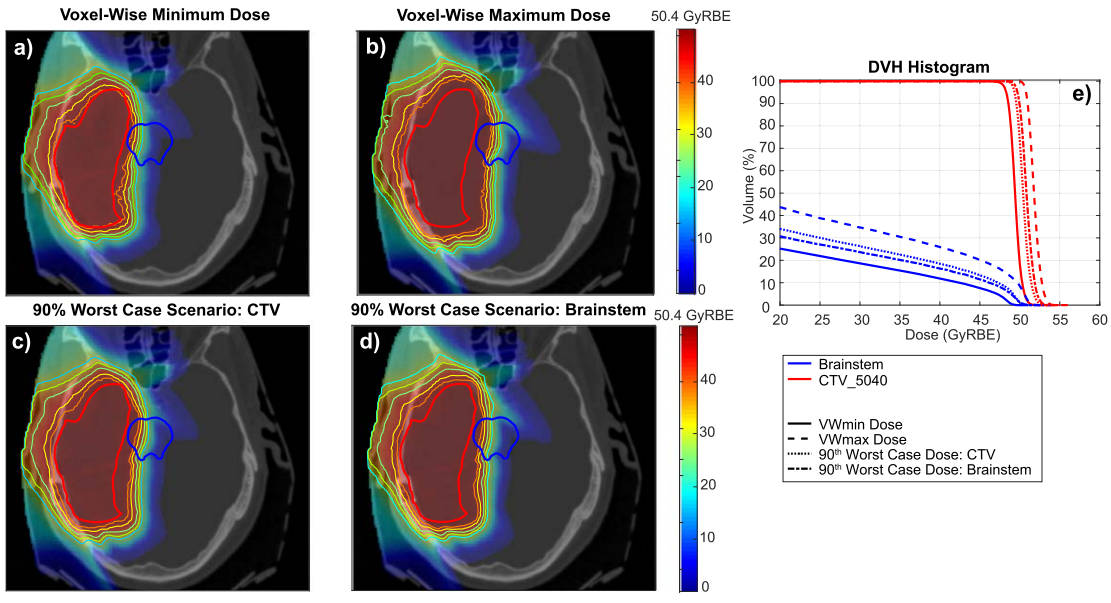

Figure 3. Comparison of the dose distributions from the VWmin scenario (A), the VWmax scenario (B) for the scenario subset in the DUPROTON protocol ($S_4 \otimes R_1$) and two extreme shifted scenarios (C and D). Extreme scenarios were chosen as a (near) worst-case scenario lying on the 90% confidence ellipsoid of the 4D error distribution for the CTV (C) and for the brainstem (D). DVH (E) for the CTV and for the brainstem are displayed in red and blue, respectively. Isodose line for the clinical constraint (orange: $D_{98\%,CTV} \geqslant 95\%\ D_{pres}$) is also shown.

3.3. Photon-based margin recipe versus consistent robustness recipes before and after protocol calibration

The robustness recipe, derived for this patient cohort from the clinical TPS, is displayed in figure 5 after (figure 5(a)) and before (figure 4(b)) calibration against the scenario subset used clinically ($S_4 \otimes R_1$). Before calibration of the protocol, the errors determined from the robustness recipe, assuming a population coverage of 90% (90% robustness recipe, blue line), are significantly larger than the errors assumed from the photon-based margin

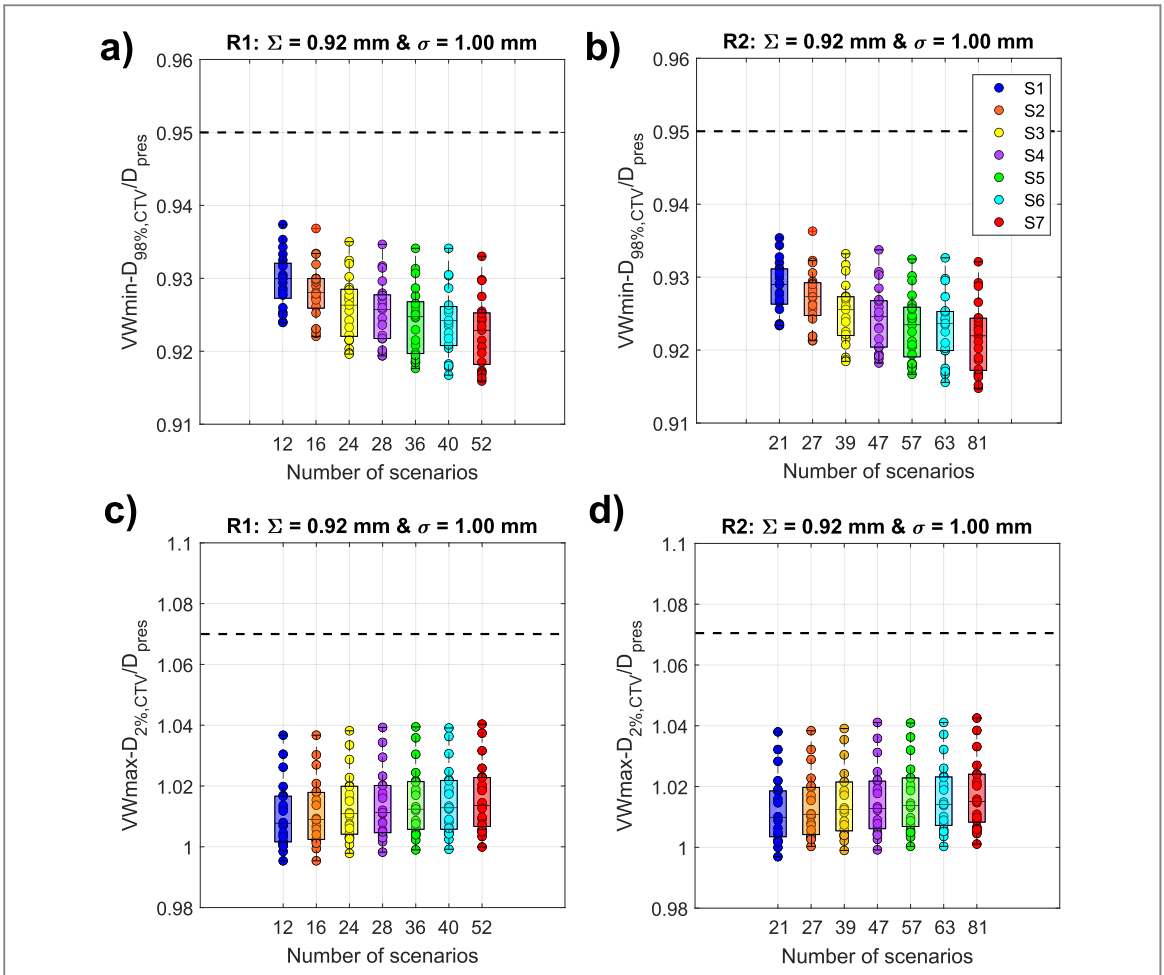

Figure 4. VWmin-$D_{98\%,CTV}/D_{pres}$ and VWmax-$D_{2\%,CTV}/D_{pres}$ dose differences among the different scenario subsets for both R_1 (left) and R_2 (right) range strategies. Subsets (S_1 to S_7) are ordered from the lowest to the largest number of geometrical error scenarios included. To reduce inter-patient variation, all values were scaled to the 10th percentile of the scenario $D_{98\%,CTV}$ distribution (90th percentile PCE-$D_{98\%,CTV} = 95\%D_{pres}$) per patient. Clinical CTV criteria are also represented as the dashed black lines for the $D_{98\%,CTV}$ ($D_{98\%,CTV} \geqslant 95\%D_{pres}$) and $D_{2\%,CTV}$ ($D_{2\%,CTV} \leqslant 107\%D_{pres}$), respectively.

recipe, also compared to its original non-linearized form. In fact, it does not reproduce the factor 2.5 that was determined in the photon-based margin recipe when the random geometrical σ error is 0 mm. For $\sigma \leqslant 1.5$ mm, both the linearized and non-linearized photon-based margin recipe could be calibrated with a linear scale factor. After calibration of the protocol, the differences between the 90% robustness recipe and the linearized and non-linearized photon-based margin recipe were reduced, but variations remained. When $\sigma > 1.5$ mm, neither form of the photon-based margin recipe can be calibrated to reproduce the robustness recipes. Small differences in the σ values lead to significant Σ differences in this part of the recipe, indicating that PT is more sensitive to random errors. For instance, a σ error = 1 mm (central part of the recipe) leads to geometrical errors of $\Sigma = 1.54$ mm (before calibration) and $\Sigma = 1.15$ mm (after calibration) according to the 90% clinical recipe, while a lower geometrical Σ error = 0.92 mm is suggested for the linearized photon-based margin recipe, which aims at the same population coverage. No fitting parameters were found for the robustness recipe aiming at a population coverage of 98% since no combination of Σ and σ errors ensured a 98% probability after calibration of the protocol. The robustness recipes for both situations (after and without protocol calibration), showed a similar consistency for different combinations of geometrical Σ and σ errors as linear photon-based margin recipes [SM, section S1]. For the robustness recipe without protocol calibration, the inter-patient variation in the scaled VWmin-$D_{98\%,CTV}$ values was higher for the different combinations of geometrical Σ and σ errors compared to the recipe after protocol calibration, indicating that the protocol might not be suitable when a large number of geometrical errors are handled in comparison to the SR setting used [SM, section S2].

4. Discussion

In this paper, we have quantitatively and systematically assessed the performance of the Dutch proton robustness evaluation protocol in a cohort of robustly planned PT treatments for 21 neuro-oncological patients. We

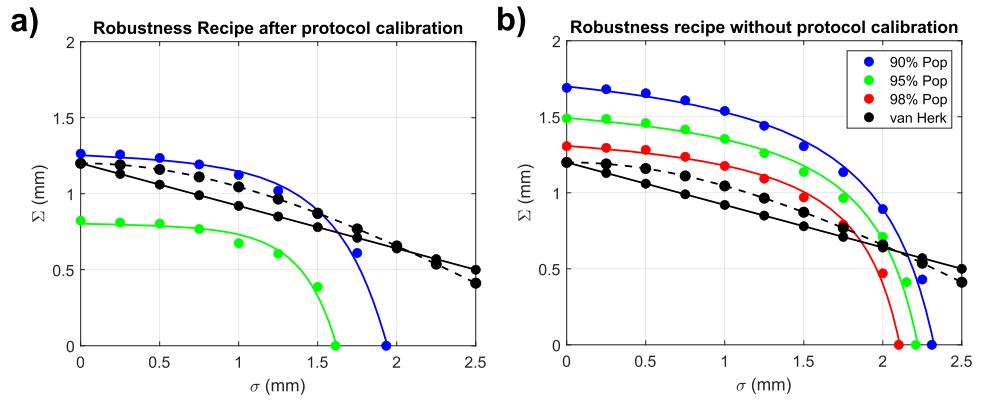

Figure 5. Robustness recipe probabilistically consistent with the TPS before (right) and after (left) the calibration of the protocol. Recipe ensures that the clinical CTV criterion ($D_{98\%,CTV} \geqslant 95\%D_{pres}$) is met at different population coverage probabilities (90% in blue, 95% in green and 98% in red). Linearized (black) and non-linearized (dashed black) photon-based margin recipes from van Herk, which aim for a 90% population coverage probability in conventional radiotherapy, are also displayed (van Herk *et al* 2004).

evaluated how VWmin and VWmax dose metrics probabilistically translate into delivered dose to the CTV and OARs under geometrical and range errors. Thus, we calibrated the DUPROTON protocol by deriving adequate CTV prescription-dose levels, assuming different scenario subsets in line with the DUPROTON group and analyzed residual inter-patient variation. Finally, a robustness recipe was determined before and after calibration of the protocol and compared to a photon-based margin recipe in which the protocol is based, to respectively assess the remaining differences when (i) SR settings are pushed to the limits to handle geometrical errors and (ii) the photon-based margin recipe is applied to PT.

The DUPROTON protocol, combined with the photon-based margin recipe to determine the adequate SR setting, can be calibrated using a lower evaluation dose level depending on the evaluation scenarios selected to construct the voxel-wise doses. In line with our findings in (Rojo-Santiago *et al* 2021a), the DUPROTON protocol ($S_4 \otimes R_1$) as implemented at our center leads to consistent but conservative results in terms of CTV and OARs doses (figure 2). Assuming a population coverage of 90%, VWmin and VWmax doses respectively result in an under- and over-estimation of 3 p.p. and 2 p.p. of the near-minimum and maximum CTV doses, respectively. The slight CTV overdose can be corrected by evaluating the VWmin CTV dose from a $L = 93.0\%$ ($S_1 \otimes R_1$) to 92.2% ($S_7 \otimes R_2$) level, instead of the usual 95%, depending on the scenario subset used for the robustness evaluation (figures 4(a) and (b)). In addition, this lower prescription-dose level does not lead to unacceptable hotspots in the delivered dose distribution. In fact, as the VWmax dose metric also overestimates the near-maximum CTV dose (figures 4(c) and (d): VWmax-$D_{2\%,CTV} = 1.02$ on average for all scenario subsets), the protocol realizes slightly more homogeneity in the delivered dose distributions compared to conventional RT plans.

In contrast, if the SR settings are pushed to the limit [SM section S1], the DUPROTON protocol is no longer conservative. In this case, the dose can be corrected by evaluating the VWmin CTV dose at a 95.6% level, which leads to more inter-patient variation in the clinical metrics. The lack of consistency of the protocol when tight robustness settings are used may be due to the limitation of using robust minimax optimization in treatment planning, which uses a discrete set of scenarios. Thus, a calibration of the protocol with probabilistic robustness evaluation approaches, which uses a semi-infinite set of scenarios, comes at the expense of increased inter-patient variation in the clinical dose metrics. In addition, the larger number of geometrical and range errors used might also contribute to the inter-patient variation, but comparable results were obtained with a cohort of head-and-neck patients planned with a SR = 5 mm setting (Rojo-Santiago *et al* 2021b). Therefore, proper probabilistic approaches for treatment plan optimization could aid in reducing the remaining inter-patient variation.

The conservativism of the DUPROTON protocol while applying photon-based margin recipes could be partially explained by (i) the inherent construction of the voxel-wise approach, in which the extreme dose levels for each voxel are reported, and (ii) the incompatibility of photon-based margins to calculate SR settings for PBS-PT, as shown in figure 5. If only Σ errors are considered, the robustness recipe does not reproduce the factor of 2.5 from the photon-based margin recipe. In addition, PT planning is more sensitive to random errors due to the steeper lateral and distal penumbrae compared to conventional RT. In fact, the remaining differences after calibration of the protocol confirm that photon-based margin recipes do not apply to PT (figure 5(b)). The differences in the degree of modulation of the intensities (conventional RT versus PBS-PT) on the treatment plans used, how the optimization was done and the assumption of a constant lateral penumbra from conventional RT and its application in PBS-PT might also contribute to these differences. Furthermore, the

point minimum dose (D_{min}) was the metric proposed to assess the plan adequacy during the construction of the photon-based margin recipe, which was used for PTV evaluation, while nowadays the near-minimum dose ($D_{98\%}$) is commonly used instead (ICRU 1993, 1999).

The VWmax-$D_{0.03cc}$ dose, which is commonly evaluated for serial OAR in clinical practice (Eekers *et al* 2018), results in a conservative metric proving that is not a good predictor of the near-maximum $D_{0.03cc}$ dose to serial organs in dose gradients. Furthermore, it depends on the dose, in which the largest absolute deviations from the unity lines in figure 2(b) are found at intermediate dose levels (relative to the prescribed dose). This is particularly relevant for cases in which robust target coverage is sacrificed to spare critical OARs, as it leads to over-estimation of the OAR dose and, therefore, to suboptimal trade-offs between target coverage and critical serial OARs. Thus, based on the DUPROTON protocol, one can give additional dose to OARs that are located close to the target if there is an improvement in CTV coverage. An example is skull-base chordomas patients, in which the prescription dose (70–74 Gy(RBE)) to the target is above critical OARs tolerances (Fung *et al* 2018, Kroesen *et al* 2022). However, a higher dosage of serial OARs should be balanced against RBE effects, which has an increased impact after the distal part of the spread-out Bragg peak (Luhr *et al* 2018).

A limitation of the study comes from the lack of knowledge about adequate probabilistic planning goals to assess target dose adequacy directly on the CTV. Clinical plan robustness evaluations depend on the robustness approaches used to mitigate uncertainties in PT (robust optimization and evaluation) and in RT (PTV-based methods), which are usually based on enlarged treated volumes around the CTV (PTV-$D_{98\%}$ for RT and VWmin-$D_{98\%,CTV}$ for PT) instead of on the CTV itself. In addition, the relaxation of the historical clinical goal from a point minimum (PTV-$D_{100\%}$) to the near-minimum dose (PTV-$D_{98\%}$) masked the volume v of the CTV that should be probabilistically covered by 95% of the D_{pres}, which has also been adopted in the DUPROTON protocol (VWmin-$D_{98\%,CTV}$). In this paper, we used the 10th percentile of the scenario $D_{98\%,CTV}$ distribution as an adequate probabilistic CTV dose metric from the PCE-based robustness evaluations, to subsequently calibrate the DUPROTON protocol. Other dose-volume metrics may be established through cross-calibration with photon treatment plans.

We limited this study to the evaluation of the clinical treatment plans in a clinical TPS, including per-patient clinical decisions and trade-offs in treatment planning, with geometrical and range robust optimization settings of 3 mm and 3%. Instead, we evaluated and optimized the performance of the protocol by using different numbers of treatment errors.

Another limitation of this study relates to the number of scenarios selected for the protocol, which were defined in line with the DUPROTON consensus group. The scenarios used in each of the subsets are highly correlated, limiting the coverage of the actual error distribution even when using a larger sample of scenarios. Thus, a more uniform sampling of the scenarios, i.e. from a fixed percentile of the 4D probability distribution used in the DUPROTON protocol might lead to a better interpretation of the clinical metrics. However, the addition of more scenarios turns the approach to a more conservative direction, as figure 4 shows. Consequently, a robustness evaluation protocol that satisfies a lower VWmin dose evaluation level and includes fewer scenarios could reduce computational time in treatment planning.

Compared to MC-based robustness evaluation methods, PCE is an analytical approximation of the dose engine that, through the computational feasibility of millions of dose calculations, can aid in accurately interpreting the impact of treatment uncertainties in fractionated treatments, in this case geometrical and SPP errors, on relevant dosimetric parameters ($D_{98\%,CTV}$, $D_{0.03cc,OARs}$) in PBS-PT. Its speed and accuracy allow us to perform probabilistic robustness evaluations, which enables us (i) to quantify the sensitivity and true robustness of these clinical dose metrics more precisely, and (ii) to benchmark other robustness strategies used in clinical practice. However, the parameterization of the treatment errors in the problem enforces a validation of the model, which might fail in the case of a more complicated source of uncertainties, i.e. anatomical variations. Furthermore, additional source of errors in the PCE construction might increase its complexity and computational cost. For other treatment sites including moving targets and anatomical deformations, the combination of PCE with more advanced anatomical modeling could further improve clinical robustness evaluation protocols (Pastor-Serrano *et al* 2021).

5. Conclusion

In summary, we have shown that the Dutch proton robustness protocol, when combined with the photon-based margin recipe to determine the adequate SR, can be calibrated with a lower VWmin evaluation dose level depending on the chosen scenario subset (e.g. $S_4 \otimes R_1$: 92.5%). PCE-based robustness evaluations showed that the protocol leads to consistent but conservative results in patients in which robustness in target coverage can be achieved. Without a dose calibration, the protocol underestimates/overestimates the near-minimum/maximum CTV doses ($D_{98\%,CTV}/D_{2\%,CTV}$) by 3 p.p./2 p.p. on average for all scenario subsets. Furthermore, in

particular, this shows limitations when assessing robustness in OAR doses. The VWmax near-maximum resulted in a poor robust metric of the near-maximum dose, especially for cases in which trade-off between robust target coverage and OAR dose must be made. Finally, the protocol might not perform well when tight SR settings are used in planning, in which the inter-patient variation in clinical dose metrics substantially increases.

Acknowledgments

This work was financiered by KWF Kanker Bestrijding (project number 11711). The authors would like to especially thank Erik Korevaar for the fruitful discussions in the development of this study and Nicola Panico for the support in the data fitting.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Jesús Rojo-Santiago https://orcid.org/0000-0002-4598-7430
Zoltán Perkó https://orcid.org/0000-0002-0975-4226

References

Bortfeld T, Paganetti H and Kooy H 2005 MOAT6B01: proton beam radiotherapy. the state of the art *Med. Phys.* **32** 2048–9
Buti G, Souris K, Barragán Montero A M, Cohilis M, Lee J A and Sterpin E 2020 Accelerated robust optimization algorithm for proton therapy treatment planning *Med. Phys.* **47** 2746–54
Eekers D B *et al* 2018 The EPTN consensus-based atlas for CT- and MR-based contouring in neuro-oncology *Radiother. Oncol.* **128** 37–43
Florijn M. A *et al* 2020 Lower doses to hippocampi and other brain structures for skull-base meningiomas with intensity modulated proton therapy compared to photon therapy *Radiother. Oncol.* **142** 147–53
Fredriksson A, Forsgren A and Hårdemark B 2011 Minimax optimization for handling range and setup uncertainties in proton therapy *Med. Phys.* **38** 1672–84
Fung V *et al* 2018 Proton beam therapy for skull base chordomas in 106 patients: a dose adaptive radiation protocol *Radiother. Oncol.* **128** 198–202
Henríquez F C and Castrillón S V 2008 A novel method for the evaluation of uncertainty in dose-volume histogram computation *Int. J. Radiat. Oncol. Biol. Phys.* **70** 1263–71
Hernandez V *et al* 2020 (www.thegreenjournal.com/article/S0167-8140(20)30813-6/fulltext)
ICRU 1993 *Prescribing, recording and reporting photon beam therapy* Report 50 (https://doi.org/10.1118/1.597396)
ICRU 1999 Report 62: prescribing, recording and reporting photon beam therapy (Suppl to 50) *Journal of the ICRU.* **74** 879
Korevaar E W *et al* 2019 Practical robustness evaluation in radiotherapy–a photon and proton-proof alternative to PTV-based plan evaluation *Radiother. Oncol.* **141** 267–74
Kosaki K *et al* 2012 Comparison of intensity modulated radiotherapy (IMRT) with intensity modulated particle therapy (IMPT) using fixed beams or an ion gantry for the treatment of patients with skull base meningiomas *Radiat. Oncol.* **7** 44
Kroesen M *et al* 2022 Single-institution clinical experience using robust intensity modulated proton therapy in chordoma and chondrosarcoma of the mobile spine and sacrum: feasibility and need for plan adaptation: Robust planning in chordoma of spine and sacrum *Radiother. Oncol.*
Langen K and Zhu M 2018 Concepts of PTV and robustness in passively scattered and pencil beam scanning proton therapy *Seminars in Radiation Oncology* **28** 248–55
Liu W *et al* 2013a Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers *Med. Phys.* **40** 1–8
Liu W, Frank S J, Li X, Li Y, Zhu R X and Mohan R 2013b PTV-based IMPT optimization incorporating planning risk volumes versus robust optimization *Med. Phys.* **40** 1–8
Lomax A 2016 *Particle Radiotherapy: Emerging Technology for Treatment of Cancer.* (Berlin: Springer) 1st edn (https://doi.org/10.1007/978-81-322-2622-2)
Lomax A J 2008a Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties *Phys. Med. Biol.* **53** 1027–42
Lomax A J 2008b Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions *Phys. Med. Biol.* **53** 1043–56
Luhr A, von Neubeck C, Krause M and Troost E G 2018 Relative biological effectiveness in proton beam therapy current knowledge and future challenges *Clin. Transl. Oncol.* **9** 35–41
Le Maître O P and Knio O M 2010 *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics* (Berlin: Springer) 1st edn (https://doi.org/10.1007/978-90-481-3520-2)
Pastor-Serrano O, Habraken S, Lathouwers D, Hoogeman M, Schaart D and Perko Z 2021 How should we model and evaluate breathing interplay effects in IMPT? *Phys. Med. Biol.* **66** 23
Perkó Z, van Der Voort S R, van De Water S, Hartman C M, Hoogeman M and Lathouwers D 2016 Fast and accurate sensitivity analysis of IMPT treatment plans using polynomial chaos expansion *Phys. Med. Biol.* **61** 4646–64

Rojo-Santiago J, Habraken S J, Lathouwers D, Méndez Romero A, Perkó Z and Hoogeman M S 2021a Accurate assessment of a Dutch practical robustness evaluation protocol in clinical PT with pencil beam scanning for neurological tumors *Radiother. Oncol.* **163** 121–7

Rojo-Santiago J, Habraken S J, Lathouwers D, Perkó Z and Hoogeman M S 2021b Limitation of van Herk's recipe in robust optimization of clinical IMPT for head and neck cancer *Abstract from the 2021 European Society for Radiotherapy and Oncology (ESTRO) meeting (Madrid, Spain)*

Sterpin E, Rivas S T, van Den Heuvel F, George B, Lee J A and Souris K 2021 Development of robustness evaluation strategies for enabling statistically consistent reporting *Phys. Med. Biol.* **66** 045002

Stroom J C, De Boer H C, Huizenga H and Visser A G 1999 Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability *Int. J. Radiat. Oncol. Biol. Phys.* **43** 905–19

Teoh S, George B, Fiorini F, Vallis K A and van den Heuvel F 2020 Assessment of robustness against setup uncertainties using probabilistic scenarios in lung cancer: a comparison of proton with photon therapy. The *Br. J. Radiol.* **93** 20190584

Unkelbach J *et al* 2018 Robust radiotherapy planning *Phys. Med. Biol.* **63** 22

Unkelbach J, Chan T C and Bortfeld T 2007 Accounting for range uncertainties in the optimization of intensity modulated proton therapy *Phys. Med. Biol.* **52** 10

Van Herk M 2004 Errors and margins in radiotherapy *Seminars in Radiation Oncology.* **14** 52–64

Van Der Voort S, van De Water S, Perkó Z, Heijmen B, Lathouwers D and Hoogeman M 2016 Robustness recipes for minimax robust optimization in intensity modulated proton therapy for oropharyngeal cancer patients *Int. J. Radiat. Oncol. Biol. Phys.* **95** 163–70

Van Dijk L V *et al* 2016 Robust intensity modulated proton therapy (IMPT) increases estimated clinical benefit in head and neck cancer patients *PLoS One* **11** 1–14

Van Der Weide H L, Kramer M C, Scandurra D, Eekers D B, Klaver Y L, Wiggenraad R G and Langendijk J A 2021 Proton therapy for selected low grade glioma patients in the Netherlands *Radiother. Oncol.* **154** 283–90

Van Herk M, Remeijer P, Rasch C and Lebesque J V 2000 The probability of correct target dosage: Dose-population histograms for deriving treatment margins in radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **47** 1121–35

Weide H L V D *et al* 2020 Proton therapy for selected low grade glioma patients in the Netherlands *Radiother. Oncol.* **154** 283–90

Wohlfahrt P, Möhler C, Richter C and Greilich S 2018 Evaluation of stopping-power prediction by dual-and single-energy computed tomography in an anthropomorphic ground-truth phantom *Int. J. Radiat. Oncol. Biol. Phys.* **100** 244–53

Wohlfahrt P, Möhler C, Stützer K, Greilich S and Richter C 2017 Dual-energy CT based proton range prediction in head and pelvic tumor patients *Radiother. Oncol.* **125** 526–33

Wohlfahrt P, Möhler C, Troost E G, Greilich S and Richter C 2019 Dual-energy computed tomography to assess intra-and inter-patient tissue variability for proton treatment planning of patients with brain tumor *Int. J. Radiat. Oncol. Biol. Phys.* **105** 504–13