Improved healthy tissue sparing in proton therapy of lung tumors using statistically sound robust optimization and evaluation

Abstract

Introduction: Robust planning is essential in proton therapy to ensure adequate treatment delivery in the presence of uncertainties. However, at both robust optimization and robust evaluation stages, commonly used techniques can be overly conservative in the way error scenarios are selected. Moreover, they typically do not provide quantified confidence levels about the robustness of the treatment. In this study, established techniques are compared to more comprehensive alternatives with the aim of evaluating the differences in target coverage and organ at risk (OAR) dose.

Method: A total of 13 lung cancer patients were planned. Two robust optimization methods were used, a conventional approach of selecting scenarios using maximum setup and range error values or selecting scenarios from marginal probabilities (SSMP) and a method that aims to address some of the statistical inconsistencies of the conventional approach by selecting errors on a predefined 90% hypersurface or scenario selection from joint probabilities (SSJP). Two robust evaluation methods were used, an approach commonly applied clinically (conventional evaluation – CE) based on generating error scenarios from pragmatic combinations of the maximum errors of each uncertainty source, and ~~the fast~~a comprehensive robustness evaluation based on the Monte Carlo dose engine MCsquare which considers scenario probabilities (statistical evaluation – SE).

Results: Nominal and worst-case scenarios were checked to assess the plan's robustness. Plans optimized using SSJP had on average 0.5 Gy lower dose in CTV D_{98} worst case than plans optimized using SSMP. When evaluated using SE, 92.3% of patients passed our clinical threshold in both optimization methods. Average gains in OAR sparing were recorded when transitioning from SSMP to SSJP: esophagus (0.6 Gy D_2 nominal, 0.9 Gy for D_2 worst case), spinal cord (3.9 Gy for D_2 nominal, 4.1 Gy D_2 worst-case), heart (1.1 Gy D_{mean}, 1.9% V_{30}), lungs- GTV (1.0 Gy D_{mean} , 1.9% V_{30}).

Conclusion: Optimization using the SSJP tool yielded significant OAR sparing in all recorded metrics (D_{mean}, V_{30}, D_2) with a target robustness within our clinical objectives provided that a more statistically sound robustness evaluation method was used, such as the SE method implemented via MCsquare.

1. Introduction

Proton therapy aims at accurately delivering curative radiation doses to tumors while reducing exposure to surrounding healthy tissue. Protons display a steep dose fall-off at the end of their range (the so-called "Bragg peak") resulting in a sharply localized dose peak. The high dose gradients, however, lead to a higher susceptibility to treatment uncertainties. Notable sources of uncertainty include, among others, setup errors, as well as range errors (stemming from the conversion of the CT Hounsfield units – HUs – to physical quantities – stopping powers). Inter- and intra- fraction motion also needs to be considered, particularly for tumors of the thorax due to breathing motion that can induce an undesirable shift or distortion of the dose distribution due to displaced density heterogeneities [1–5]. Given this, taking uncertainties into account in the planning stage is of paramount importance. This can be achieved via robust optimization which directly incorporates treatment errors in the optimization process [6–9].

One of the most widely used robust optimization methods, known as 'worst-case' robust optimization, aims at achieving adequate target coverage by using combinations of treatment errors to generate scenarios [8,10]. In popular implementations of worst-case robust optimization, such as Fredriksson's "minimax" optimization [8], scenarios are evaluated after each iteration during the optimization process ensuring that the objective function of the current worst-case scenario is minimized. In typical clinical implementations of the worst-case robust optimization workflow, several issues can be identified. Firstly, overly conservative scenarios are being pre-selected due to the scenarios being composed of maximum errors of each uncertainty source [8]. For a lung tumor case, the following are commonly used: ± 5 mm setup error in each direction, ± 3% image conversion error and three breathing phases including the maximum inhale and exhale [11,12]. As mentioned by Korevaar et al [13] and by Sterpin et al [14] this amounts to extremes of marginal probability distributions being combined instead of sampling the joint probability distributions. Secondly, a lack of consistently calculated confidence levels leads to the concept of a 'worst-case' becoming hard to define [14]. Said limitations are typically met both in robust optimization and evaluation.

The marginal approach is used in good clinical practice both in optimization and evaluation. Alternative approaches have been suggested in literature that aim to improve upon it. For robust optimization, Buti et al developed a method of preselecting a set of treatment error scenarios by considering the systematic setup and range uncertainties' joint probabilities [15]. Korevaar et al [13] performed robust evaluation using a statistically consistent but limited set of scenarios. Robustness evaluation with MCsquare, a Monte Carlo dose engine developed by Souris et al [14,16], enables exploring the dosimetric error space in a more statistically consistent manner at a 90% confidence level. Another approach for performing a comprehensive evaluation of a plan's robustness is the polynomial chaos expansion method as described by Perkó et al [17] which can accurately estimate the dose, its variance and distribution in any particular error scenario. To address the loss of robustness in lung cases with significant motion, Taasti et al [18] proposed a joint treatment planning and robust evaluation approach based on generating an internal target volume (ITV) achieving clinically viable plans.

In this publication, we would like to explore whether a clinical benefit can be expected using scenario selection tools with improved statistical foundations, both at the level of robust optimization and evaluation. A workflow including worst-case robust optimization and evaluation via RayStation as performed conventionally in clinical practice is compared to two other tools: a tool that enables scenario selection from joint probabilities developed by Buti et al [15] and MCsquare [16]. We have chosen here lung tumors, because of the challenges this location entails with respect to robust planning. By applying those methods on realistic clinical cases, we aim at evaluating their impact on target coverage and organs-at-risk sparing.

2. Material and Methods

2.1. Patient data

The planning database contained 13 lung cancer patients. Patient data consisted of a 4D-CT image set containing ten, evenly spaced in time, breathing phases. The 13 patients' data has been used retrospectively in previous studies. A 60 Gy dose prescription over 30 fractions was used with a CTV coverage goal of delivering at least 95% of the prescribed dose (= 57 Gy) to 98% of the target volume. Constraints were placed on the organs-at-risk (OARs) on a case-by-case basis, depending on tumor and lymph node size and positioning and proximity to OARs. In all cases, priority was given to

maintaining target coverage while remaining below the OAR constraints in the Appendix (Table S3). Only the target was robust optimized. The CTV size and position relative to the lung for all patients are given in the Appendix (Table S4).

All treatment plans used the MidP-CT as the nominal planning CT [19]. For robust optimization, three additional breathing phases were used: the maximum exhale CT (End_ExH), the maximum inhale CT (End_InH) and the mid ventilation CT (MidV) [20]. When evaluated all ten breathing phases were used.

2.2. Robust optimization

Robust optimization was performed on RayStation 9B on a computer with the following specifications: Intel Xeon Gold 6234 CPU (two 3.30 GHz processors), 128 GB RAM, NVIDIA Quadro RTX 6000 GPU, 2 TB SSD.

Two worst-case scenario selection methods were compared: conventional scenario selection from marginal probabilities (SSMP) and scenario selection from joint probabilities (SSJP), a method of preselecting a limited set of treatment error scenarios developed by Buti et al [15]. Scenarios cover geometric uncertainties (setup errors, range errors, motion), interplay was not considered for this study.

2.2.1. SSMP

In SSMP, maximum setup errors calculated using systematic (Σ) and random (σ) setup and baseline shift values were used for two cases: tumor only and tumor with lymph nodes. This was done using van Herk's margin formula [21] with the goal of obtaining a margin that ensures a minimum dose is delivered to 90% of the patient population:

$$M = 2.5 \cdot \Sigma_{total} + 0.7 \cdot \sigma_{total} \quad (1)$$

The limitations of such a simplified approach for determining margins in proton therapy must be acknowledged. Strictly speaking, simple margin recipes cannot be derived in a sound statistical manner in proton therapy because of the failure of the static dose cloud approximation [14]. However, as observed by Korevaar et al [13] when attempting to provide a practical, PTV-less approach for proton treatment planning, applying van Herk's formula (1) when converting uncertainties into errors in robust optimization is a limited yet suitable approach in most cases.

The calculated total setup error values for each direction were inputted in our treatment planning system (TPS), RayStation. However, using a margin that exceeds 5 mm in RayStation leads to a significant increase in optimization time due to the system automatically generating intermediate errors to ensure robust coverage, which can be an overconservative approach. As a workaround, instead of optimizing on the CTV, a patient-specific CTV expansion can be generated. From the full value of the margin, 5 mm is subtracted (in each direction) and subsequently used as an isotropic setup error value. The CTV is then expanded by the remaining amount. The complete setup margin values can be found in the Appendix (Table S1 and Table S2). Optimization was done on the expanded CTV volume to reduce computation time, while evaluation was done on the CTV using the total uncertainty values.

The standard deviation of the range error resulting from the conversion of the CT Hounsfield units (HUs) to relative stopping power was set equal to 2.6% for optimization and 1.6% for evaluation as

per reviewing Paganetti et. al [2]. The total number of optimization scenarios is 63: 7 (setup errors: ± 5 mm in x,y,z directions, additionally the nominal scenario) × 3 (image conversion errors: ±3%, 0%) × 3 (breathing phases: MidP, maximum inhale and maximum exhale).

2.2.2. SSJP

The SSJP method aims to address some of the statistical inconsistencies of the conventional approach. Details on the SSJP method can be found in Buti (2019) [15]. In short, by considering the systematic setup and range uncertainties' joint probabilities, a 90% 4D-equiprobability hypersurface can be defined as seen in Figure 1. Twelve scenarios that do not exceed the maximum systematic setup error given by $2.5 \cdot \Sigma$ are selected on the hypersurface. For these scenarios, an additional error of $0.7 \cdot \sigma_{total}$ is added to the systematic setup error in order to obtain errors of magnitude as given by equation (1). Seventeen additional scenarios are selected in the whole hypervolume (excluding values higher than $2.5 \cdot \Sigma$) that cover any estimated residual range errors (both under and overshoot). To include these scenarios, first, proton ranges are estimated by converting the breathing CTs into maps of water-equivalent path lengths (WEPLs). Because WEPLs are beam specific, each breathing phase will have a separate WEPL map for each beam angle. The WEPL values are scaled with the range error value leading to a distribution of WEPL values for all target voxels, across all scenarios. The number of voxels that a scenario has in common with the minimum or maximum WEPL allows for the worst case over- and undershoot to be identified.

Considering the nominal scenario, this sums up to a total of 30 scenarios. For each scenario, a virtual CT is generated that represents the selected error. These virtual CTs can be imported in our treatment planning system (TPS) and selected as the set of error scenarios.

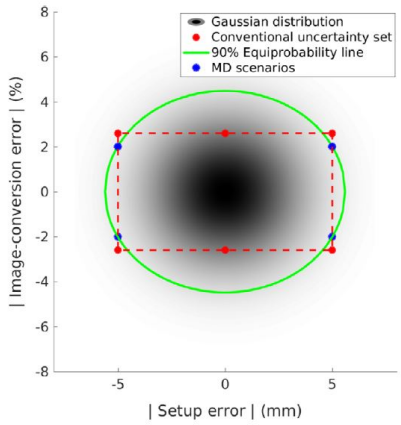

Figure 1. Two-dimensional projection of a 4D-Gaussian probability distribution representing the likelihood of sampled scenarios (the lighter, the more unlikely). The 90% equiprobability line (green) defines all possible scenarios that are positioned exactly on the edge of the 90% confidence interval. The scenarios within the conventional uncertainty set (combinations of ±5 mm setup errors and flat ±2.6% image-conversion errors) are depicted by the red circles. The maximum displacement (MD) scenarios are depicted by the blue circles (4 scenarios in 2D). Taken from Buti et al [15].

2.3. Robustness evaluation

The plans were evaluated using two methods: a conventional RayStation approach via scripting that uses setup and range errors to calculate perturbed dose scenarios or conventional evaluation (CE) and a more statistically sound method implemented in MCsquare, a Monte Carlo dose engine developed by Souris et al [16] to realistically simulate proton PBS treatments or statistical evaluation (SE).

CE consists of combinations of the maximum errors of each uncertainty source such as setup, image conversion and three breathing phases (maximum inhale and exhale and mid ventilation). Similarly to the SSMP robust optimization case, 63 evaluation scenarios are used: 7 (setup errors: ± 5 mm in x,y,z directions, additionally the nominal scenario) × 3 (image conversion errors: ±3%, 0%) × 3 (breathing phases: MidP, maximum inhale and maximum exhale).

SE takes a more comprehensive approach by randomly sampling error scenarios and recomputing the dose distributions for all of them while discarding the 10% worst scenarios based on the target D_{95} [22]. Each simulated scenario represents the entirety of the treatment (30 fractions comprising 10 breathing phases each). This contrasts the conventional approach in which a scenario represents only a combination of errors occurring in a specific breathing phase, for a specific fraction setup because of the lack of correctly modeled random errors. Each SE fraction is individually simulated considering systematic errors, sampled once per full treatment simulation and random errors re-sampled for each fraction [22]. SE models setup errors by shifting the beam isocenter and the range errors by scaling CT densities [22]. By defaults, breathing motion is simulated by recomputing the dose distribution for each breathing phase and accumulating the dose on the mid-position CT (MidP-CT), after non-rigid registration of each breathing phase to the reference phase [22], and referred in the rest of this manuscript as SE. In this work, we also consider an ITV-like approach in which random and systematic errors were compiled for each breathing phase and the worst-case was determined out of the 10 computed dose maps (SE_ITV). SE_ITV is used with the goal of ensuring the ITV is covered by the minimum clinical dose threshold at every phase, acting as a target-coverage control method.

Overall a complete SE evaluation is composed of 100 scenarios. The number of scenarios was chosen due to the convergence of the D_{95} uncertainty bands from 100 scenarios onwards maintaining a balance between its confidence interval and impact of the statistical noise and computation time for our patient data size [22]. A 1.17 × 1.17 × 2 dose grid resolution and 10^8 ions per spot were used. The evaluation results come in the form of dose-volume histograms (DVHs). During evaluation we report for the target coverage the D_{98} (Gy) nominal and worst case values as well as D_{mean} (Gy) and V_{30} (%) for heart and lungs-GTV and D_2 (Gy) for spinal cord and esophagus.

2.4. Statistical analysis

A paired t-test was performed to assess the statistical significance of the differences in target and OAR dose between our robust optimization methods and between our evaluation methods. The t value is calculated by using formula (2):

$$T = \frac{m_1 - m_2}{\sigma_d/\sqrt{n}} * 100 \qquad (2)$$

Where m_1 and m_2 are the mean values of each sample set, σ_d is the standard deviation of the differences of the paired data values and n is the sample size or the number of paired differences. The

t value represents the ratio between the difference between the two sets of data and the difference within them, as such a lower t value correlates to the two data sets being more similar. The *p* value is also automatically calculated by using the sampling distribution of the test statistic under the null hypothesis with a 95% confidence interval or 5% alpha level, meaning that a value lower than 0.05 or 5% indicates the data did not occur by chance.

The mean dose difference as a percentage of the maximum clinical dose constraint between the two methods was calculated as seen below in formula (3):

$$\Delta D = \frac{\bar{d}}{M} * 100 \quad (3)$$

$\bar{d}$ is the mean difference between dose metrics and M is the organ-specific maximum dose (30 Gy for esophagus, 20 Gy for heart, spinal cord and lungs-GTV, 60 Gy for the CTV).

3. Results

The results obtained for CTV coverage quantified by D_{98} in the nominal and worst-case for each scenario selection and robust evaluation method combination are provided in Table 1 and Table 2. An example of individual DVH metrics for patient 9 is shown in figure 2. Table 3 provides the results comparing SE and SE_ITV, verifying that the dose distributions meet our target coverage criteria without significant deviations. In terms of the OAR, the average gains in OAR sparing quantified by D_{mean}, D_2 and V_{30} when going from SSMP to SSJP are provided in table 4. The OAR sparing data is provided in the Appendix (Table S5) as well as the complete data for SSMP and SSJP (Table S6 and S7).

Table 1. D_{98} (Gy) target values for evaluated patients. Nominal values represented in black and worst-case values represented in red. SSMP – robust optimization with scenario selection from marginal probabilities, SSJP – robust optimization with scenario selection from joint probabilities, CE – conventional evaluation, SE – statistical evaluation.

Patient	SSMP		SSJP	
	CE	SE	CE	SE
1	58.9/57.2	59.4/58.8	58.9/55.4	59.4/58.8
2	58.6/56.5	58.9/56.6	58.5/55.1	58.9/56.4
3	58.7/57.1	59.1/58.7	58.7/56.7	59.2/57.8
4	58.6/57.3	58.9/58.0	58.6/56.2	59.1/57.3
5	58.8/57.8	59.4/58.8	58.8/57.6	59.4/58.6
6	58.7/57.6	59.4/58.7	58.6/57.2	59.4/58.0
7	58.5/56.6	59.3/57.9	58.6/56.8	59.3/57.5
8	58.8/57.8	59.0/58.3	58.8/58.0	59.1/57.5
9	58.8/57.9	59.4/58.9	58.9/57.2	59.5/58.8
10	58.9/57.3	59.6/59.1	58.7/57.5	59.7/58.6
11	58.6/57.7	59.2/59.0	58.6/57.0	59.3/58.7
12	58.8/56.9	59.5/58.9	58.8/56.0	59.5/58.9

Commenté [ES1]: In all table and figure captions, I would add this for SSMP and SSJP as I for Table 1

13	58.9/57.5	59.3/58.8	58.8/56.1	59.3/58.7

Table 2. Average nominal and worst case D_{98} (Gy) for each scenario selection and evaluation method combination used. SSMP – robust optimization with scenario selection from marginal probabilities, SSJP – robust optimization with scenario selection from joint probabilities, , CE – conventional evaluation, SE – statistical evaluation.

Scenario selection method	Evaluation method	D_{98} average nominal [Gy]	D_{98} average worst-case [Gy]
SSMP	CE	58.7 ± 0.1	57.3 ± 0.4
	SE	59.3 ± 0.2	58.5 ± 0.7
SSJP	CE	58.7 ± 0.1	56.7 ± 0.8
	SE	59.3 ± 0.2	58.1 ± 0.7

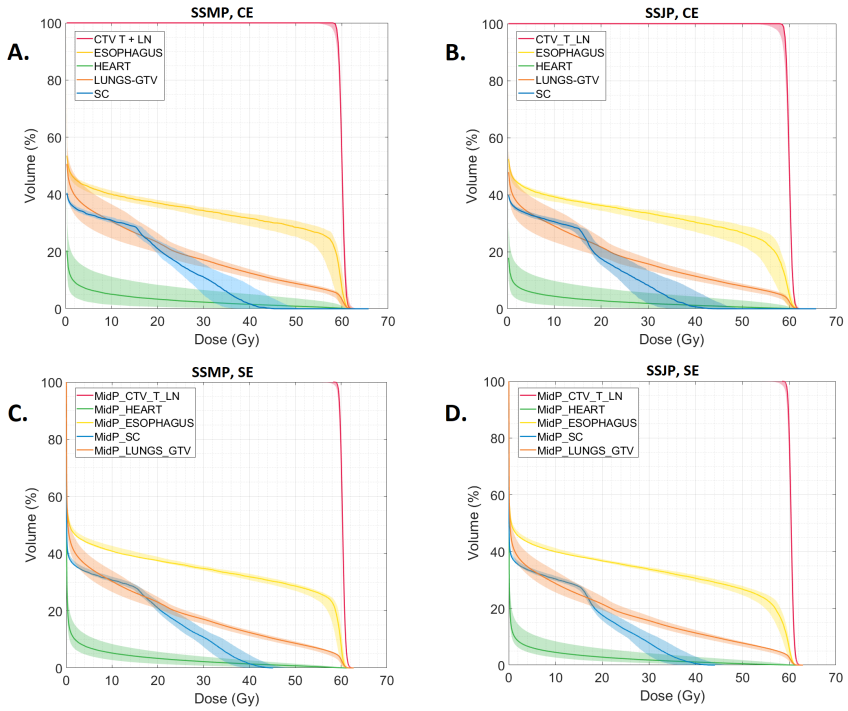

Figure 2. DVH results of the robustness evaluation for patient #9: A – scenario selection from marginal probabilities and conventional evaluation, B - scenario selection from joint probabilities and conventional evaluation, C – scenario selection from marginal probabilities and statistical evaluation, D - scenario selection from joint probabilities and statistical evaluation. Note the narrowing of the bands when evaluating with SE.

Table 3. Comparison of D_{98} (Gy) target values between SE which simulates breathing by recomputing the dose distribution for each phase and accumulating the dose on the MidP-CT and SE_ITV which computes dose maps for each breathing phase. SSMP - scenario selection from marginal probabilities, SSJP - scenario selection from joint probabilities. Nominal values represented in black and worst-case values represented in red.

Patient	SSMP		SSJP	
	SE	SE_ITV	SE	SE_ITV
1	59.4/58.8	59.0/58.7	59.4/58.8	59.0/58.8
2	58.9/56.6	58.2/56.7	58.9/56.4	58.4/56.4
3	59.1/58.7	58.7/58.3	59.2/57.8	58.9/57.6
4	58.9/58.0	58.5/58.0	59.1/57.3	58.7/57.1
5	59.4/58.8	58.9/58.5	59.4/58.6	58.9/58.5
6	59.4/58.7	59.0/58.3	59.4/58.0	59.0/58.0
7	59.3/57.9	58.8/57.8	59.3/57.5	58.9/57.4
8	59.0/58.3	58.7/57.1	59.1/57.5	58.8/58.1
9	59.4/58.9	59.0/58.7	59.5/58.8	59.1/58.1
10	59.6/59.1	59.2/58.9	59.7/58.6	59.2/58.8
11	58.6/57.7	58.6/58.7	58.6/57.0	58.6/58.3
12	59.5/58.9	59.5/58.7	59.5/58.9	59.6/58.6
13	59.3/58.8	59.2/58.3	59.3/58.7	58.8/58.1

The SSJP tool showed lower levels of target robustness than SSMP plans, as visualized in Figure 3. On average, SSMP optimized plans had a 0.5 Gy higher dose in the D_{98} worst-case as opposed to their SSJP counterparts. In four out of the thirteen cases, switching from SSMP to optimizing with SSJP while CE led to the worst-case scenario dose falling below our predefined 57 Gy threshold. Three of the patients did not have adequate target coverage using SSMP and CE which remained the case after being optimized with our alternative tool. This was, however, not the case when evaluated with SE.

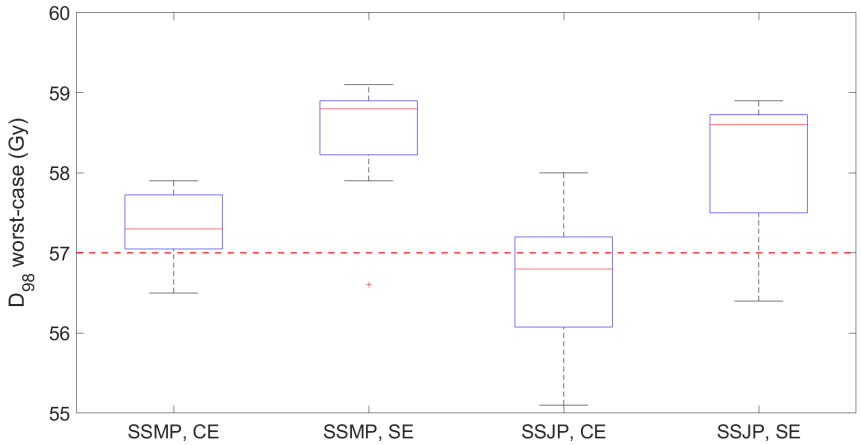

Figure 3. CTV D_{98} (Gy) target values with the minimum dose threshold highlighted with the dashed red line. The blue boxes represent the interquartile ranges (IQR) and the red lines inside them the median values. Whiskers determined by the furthest value in the interval between the 25th percentile minus 1.5*IQR and the 75th percentile plus 1.5*IQR. Outliers marked in red. SSMP - scenario selection from marginal probabilities, SSJP - scenario selection from joint probabilities, CE – conventional evaluation, SE – statistical evaluation.

Gains in terms of OAR sparing can be seen when optimizing with the SSJP tool in all recorded metrics: up to 1.8 Gy in heart D_{mean}, 2 Gy in lung-GTV D_{mean}, 9.7 Gy in spinal cord D_2 nominal, 8.8 Gy in spinal cord D_2 worst case, 3.1 Gy in esophagus D_2 nominal and 3.6 in esophagus D_2 worst case. The OAR results for each metric are visualized in Figure 4, Figure 5, and Figure 6 with the complete data in Appendix (Table S5).

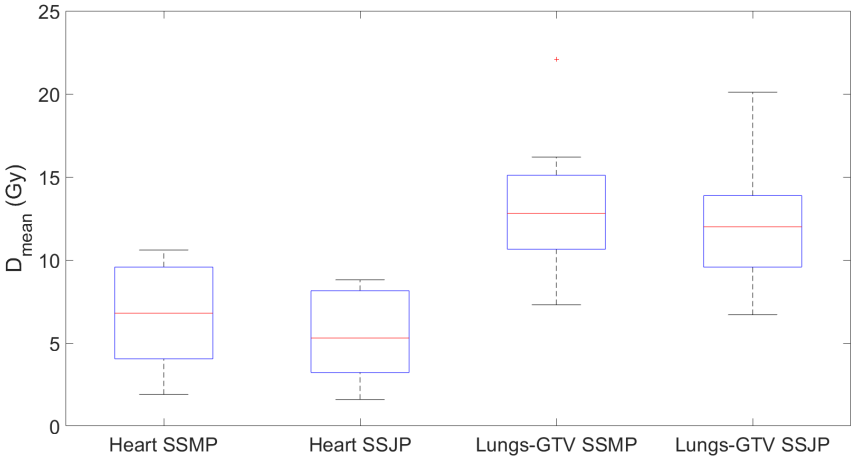

Figure 4. D_{mean} (Gy) for heart and lungs-GTV, conventionally evaluated. The blue boxes represent the interquartile ranges (IQR) and the red lines inside them the median values. Whiskers determined by the furthest value in the interval between the 25th percentile minus 1.5*IQR and the 75th percentile plus 1.5*IQR. Outliers marked in red. SSMP - scenario selection from marginal probabilities, SSJP – scenario selection from joint probabilities.

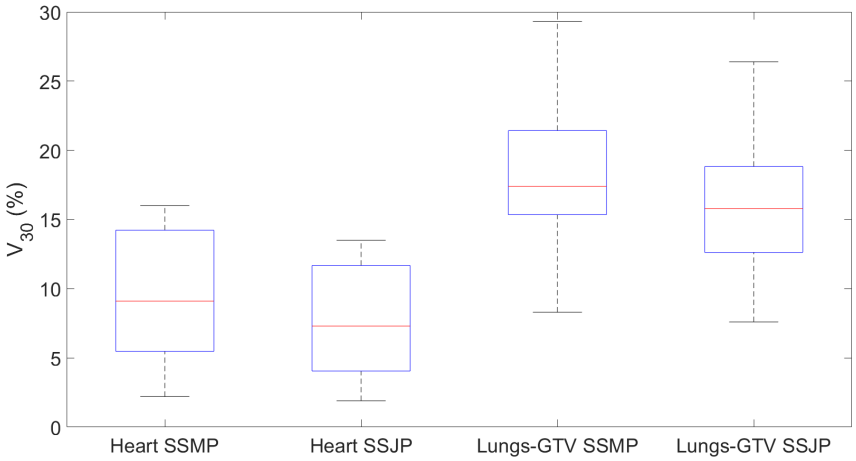

Figure 5. V_{30} (%) for heart and lungs-GTV, conventionally evaluated. The blue boxes represent the interquartile ranges (IQR) and the red lines inside them the median values. Whiskers determined by the furthest value in the interval between the 25th percentile minus 1.5*IQR and the 75th percentile plus 1.5*IQR. Outliers marked in red. SSMP - scenario selection from marginal probabilities, SSJP - scenario selection from joint probabilities.

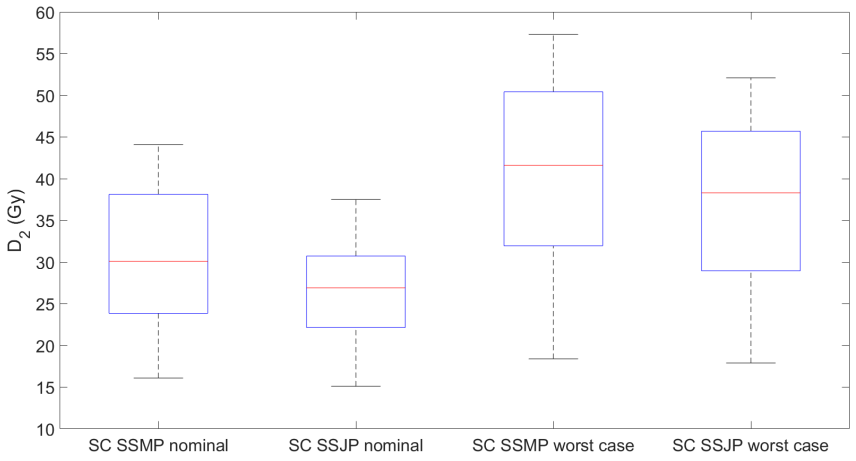

Figure 6. D_2 (Gy) for spinal cord (SC) nominal and worst case, conventionally evaluated. The blue boxes represent the interquartile ranges (IQR) and the red lines inside them the median values Whiskers

determined by the furthest value in the interval between the 25th percentile minus 1.5*IQR and the 75th percentile plus 1.5*IQR. Outliers marked in red. SSMP - scenario selection from marginal probabilities, SSJP - scenario selection from joint probabilities.

Table 4. Average gains in D_{mean} (Gy), $D_{2\ nom}$ nominal (Gy), $D_{2\ wc}$ worst case (Gy) and V_{30} (%) when switching from scenario selection from marginal probabilities (SSMP) to scenario selection from joint probabilities (SSJP). Evaluated conventionally (CE) as seen in section 2.3.

OAR	$\Delta D_{2\ nom}$ (Gy)	$\Delta D_{2\ wc}$ (Gy)
Esophagus	0.6	0.9
Spinal cord	3.9	4.1
	D_{mean} (Gy)	V_{30} (%)
Heart	1.1	1.9
Lungs-GTV	1.0	1.9

When conventionally evaluated (CE), three SSMP and seven SSJP patients did not meet our minimum worst-case scenario threshold. On the other hand, when the evaluation is performed by SE, only patient #2 failed to meet the clinical threshold for both methods. D_{98} target values did not vary significantly between SE and SE_ITV as seen in Table 3.

To verify the statistical relevance of the difference between the sets of data, results from a paired t-test in the form of the t value, the p value and the mean dose difference as a percentage of the organ-specific maximum dose can be seen below in Table 5 for OAR and Table 6 for the target.

Table 5. T value, p value and mean dose difference as a percentage of the maximum clinical dose constraint (ΔD) between SSMP and SSJP optimization OAR data. Evaluated conventionally (CE) as seen in section 2.3.

D_{mean} (Gy)	T-value	P-value	ΔD (%)
Heart	9.2	$8.5 \cdot 10^{-7}$	5.4
Lungs-GTV	6.6	$2.5 \cdot 10^{-5}$	5.2
	D_2 (Gy)		
Spinal cord (nominal)	3.9	$0.3 \cdot 10^{-2}$	19.4
Spinal cord (worst case)	6.2	$2.0 \cdot 10^{-2}$	20.4
Esophagus (nominal)	2.7	$2.0 \cdot 10^{-2}$	3.0
Esophagus (worst case)	3.3	$0.7 \cdot 10^{-2}$	4.3
	V_{30} (%)		
Heart	6.8	$2.0 \cdot 10^{-5}$	-

Lungs-GTV	5.9	$6.8 \cdot 10^{-6}$	-

Table 6. T value, p value and mean dose difference as a percentage of the target dose prescription (ΔD) for worst-case. SSMP - scenario selection from marginal probabilities, SSJP - scenario selection from joint probabilities, CE – conventional evaluation, SE – statistical evaluation.

Method 1	Method 2	T-value	P-value	ΔD (%)
SSMP, CE	SSMP, SE	-8.0	$3.8 \cdot 10^{-6}$	-2.0
SSJP, CE	SSJP, SE	-5.1	$2.8 \cdot 10^{-4}$	-2.4
SSMP, SE	SSMP, SE_ITV	3.3	$6.4 \cdot 10^{-3}$	0.5
SSJP, SE	SSJP, SE_ITV	1.5	0.2	0.2

4. Discussion

We aim to assess the individual impact of each robust optimization (SSMP and SSJP) and evaluation method (CE and SE) as well as their combinations to establish an optimal planning strategy. By comparing said combinations in terms of dose metrics, their impact on target coverage and OAR sparing can be seen. The benefits of the increased healthy tissue sparing can further be discussed regarding each method's clinical viability.

Statistically and clinically significant gains in terms of OAR sparing were recorded for all metrics (D_{mean},D_2,V_{30}) when using SSJP as seen in Figure 4, 5 and 6. Except for one patient, we have not observed clinically significant differences for the esophagus D_2, due to it being completely within or having a significant portion of its volume within the target ITV in patients 2-13. Comparing both scenario selection methods, plans optimized with the SSJP tool showed lower levels of target robustness than SSMP plans. This was expected as the SSJP tool aims at securing robustness at a predefined 90% confidence level with the aim of achieving a level of target robustness situated at the limit of clinical acceptability [15]. Most robust treatment strategies found in literature select setup and range errors separately, without considering confidence levels. However, the correlation between an increase in target robustness and a higher dose to OAR is made apparent in both worst-case "minimax" optimization and the PTV approach [23–25]. Conventional treatment planning tends towards inputting a series of constraints and seeing whether the plan's robustness is on par with our target coverage criteria. Ideally when using SSJP, a robust objective, namely securing our target coverage at the limit of clinical acceptability, is considered by default. The reduction of the target margin to the bare minimum is the main drive that enables substantial and consistent OAR sparing results, further highlighting the need for an alternative robust evaluation approach that provides an estimate of robustness at the limit of clinical acceptability (i.e. adequate coverage for at least 90% of patients).

The choice of the evaluation method did impact whether the minimum target coverage threshold was achieved. The CE approach that generates error scenarios based on combinations of the maximum errors of each uncertainty source (setup, range and breathing) led to three SSMP and seven SSJP patients not meeting our minimum worst-case scenario threshold. This results from the selection of extreme scenarios outside the 90% hypersphere in the scenario space, therefore beyond the commonly accepted 90% confidence level. As per Van Herk [21] when using a PTV margin approach,

in order to assure target robustness, appropriate confidence levels need to be established, quantifying to at least 90%. SE allows for a better definition of the confidence interval due to it evaluating in the dosimetric space instead of the scenario space.

When using SE and sampling scenarios from the entire dosimetric space [22] only patient #2 failed to meet our clinical threshold for both scenario selection methods. This can be attributed to patient #2 having a distinct layout in terms of the number of targeted lymph nodes and their distribution across both lungs which makes compromising between meeting our target goals and not overdosing the OARs significantly more difficult. Results obtained using SE can be attributed to the more detailed, better quantified, and less conservative way it operates. By randomly sampling error scenarios and recomputing the dose distribution for the best 90% scenarios based on target D_{95} it overall offers a more realistic view on the robustness of the treatment than its conventional counterpart. This discrepancy has significant clinical implications. A more conservative evaluation approach such as the commonly-used worst-case conventional evaluation (CE) has the potential of leading to results that do not pass our clinical criteria since the method by which scenarios are selected only takes into account the maximum errors of each source, not the joint probability of these magnitudes occurring during the treatment process. In clinical practice, this can lead to replanning due to the lack of target coverage, further potentially leading to an increase in OAR dose. Use of a more statistically sound robust evaluation method, such as SE, has the potential to save time within the treatment workflow by reducing the need for further optimization. It should be noted that our approach was limited by both the lack of consideration for the interplay effect as well as the patient-specific CTV expansion done in an effort to reduce treatment planning complexity, as a workaround for having a setup error larger than 5 mm.

Our planning objective was ensuring the CTV was covered in all phases, given this, it is intuitive to evaluate our target's ITV coverage when using SE. However, this did not properly correlate to how breathing motion was simulated in our initial use of SE. For SE breathing is simulated by recomputing the dose for each breathing phase and accumulating the dose on the mid-position CT (MidP-CT). To ensure CTV coverage in all breathing phases, a second, control SE approach was used. When using SE_ITV, systematic errors were compiled for each breathing phase and the worst-case was determined out of the 10 computed dose maps. Comparing SE and SE_ITV methods, negligible differences were noted as seen in table 3, which confirms the CTV is covered by the minimum clinical dose threshold at every phase. SE can be safely implemented as an evaluation tool leading to SSJP becoming a viable scenario selection option for improved OAR sparing while maintaining acceptable levels of target robustness.

5. Conclusions

Establishing a proper robust optimization and evaluation workflow is essential to realize the potential of proton therapy. Choosing the appropriate methods is both a matter of considering their statistical consistency as well as pragmatic factors such as processing time and whether the emphasis should be placed on maintaining target coverage or sparing adjacent OARs on a patient-by-patient basis.

Two methods of selecting treatment scenarios for robust optimization were used and compared: scenario selection from marginal probabilities (SSMP) and scenario selection from joint probabilities (SSJP), a method of preselecting a statistically-sound set of treatment error scenarios developed by

Buti et al [15]. For evaluating the robustness of the plans, the conventional approach (CE) and statistical evaluation (SE) were used.

Use of the SSJP tool led to significant OAR sparing in all recorded metrics (D_{mean}, V_{30}, D_2) with a target robustness within our clinical objectives provided that a statistically sound and comprehensive robustness evaluation method was used (SE). This highlights the importance of using both advanced optimization and evaluation tools when we aim at ensuring a quantified level of robustness.

Acknowledgements

Vlad Mihai Badiu is supported by Stichting tegen Kanker (Grant reference number FAF-C/2018/1195). Kevin Souris is funded by the Walloon region (MECATECH/BIOWIN, grant number 8090).

Appendix. Additional Figures/Tables

Table S1. Setup margins for tumor only case.

	$X_{(sagittal)}$ (mm)	$Y_{(coronal)}$ (mm)	$Z_{(transverse)}$ (mm)
Σ_{BL}	1.8	1.6	1.9
Σ_S	1.6	2	2.4
σ_{BL}	1.6	1.6	2.1
σ_S	1.8	2.1	2.1
Σ_{total}	2.41	2.56	3.06
σ_{total}	2.41	2.64	2.97
M_{PTV}	7.71	8.25	9.73
Expanded CTV margin	2.71	3.25	4.73

Table S2. Setup margins for tumor + lymph node.

	$X_{(sagittal)}$ (mm)	$Y_{(coronal)}$ (mm)	$Z_{(transverse)}$ (mm)
Σ_{BL}	1.9	1.6	1.9
Σ_S	1.6	2	2.4
σ_{BL}	1.7	1.6	2.1
σ_S	1.8	2.1	2.1
Σ_{total}	2.48	2.56	3.06
σ_{total}	2.48	2.64	2.97
M_{PTV}	7.94	8.25	9.73
Expanded CTV margin	2.79	3.25	4.73

Table S3. OAR constraints.

OAR	$D_{0.035cc}$ (Gy)	D_{mean} (Gy)	V_{30} (%)
Esophagus	< 60	-	-
Heart	< 63	< 20	-
Lungs-GTV	-	< 20	< 20
Spinal cord	< 50	-	-

Table S4. CTV volume (cm^3) and location relative to the lungs.

Patient	CTV (cm^3)	CTV location
1	146	Right superior lobe
2	149	Left superior lobe
3	316	Right superior lobe
4	136	Right superior lobe
5	165	Left superior lobe
6	76	Right superior lobe
7	119	Right middle lobe
8	400	Left superior lobe
9	119	Left superior lobe
10	119	Right superior lobe
11	300	Right middle lobe
12	214	Right superior lobe
13	268	Right superior lobe

Table S5. Gains in D_{mean} (Gy), D_2 nominal (Gy) and V_{30} (%) when switching from scenario selection from marginal probabilities (SSMP) to scenario sele from joint probabilities (SSJP). Evaluated conventionally (CE) as seen in section 2.3.

OAR	Unit	Patient												
		1	2	3	4	5	6	7	8	9	10	11	12	1
Esophagus	$D_{2\,nom}$	0.8	0.5	-0.2	0.4	0.3	3.1	0.2	0.3	0.2	0.5	0.2	0.8	0.
	$D_{2\,wc}$	3.6	1.0	1.7	0.6	1.1	0.6	1.0	0.1	0.3	-0.1	0.3	0.7	0.
Heart	D_{mean}	0.6	1.8	1.3	1.5	1.5	0.9	1.2	1.2	0.3	0.6	1.2	0.8	1.
	V_{30}	0.9	4.4	2.3	2.5	2.7	1.5	1.9	2.1	0.3	1.0	1.8	1.4	1.
Lungs-GTV	D_{mean}	0.4	2.0	1.3	1.7	1.9	1.0	1.2	1.1	0.7	0.6	0.8	0.8	0.
	V_{30}	0.7	4.1	2.3	3.1	3.2	1.3	2.7	1.7	1.4	0.9	1.6	1.3	0.
Spinal cord	$D_{2\,nom}$	1.0	0.2	9.7	7.6	1.9	8.0	2.1	6.9	1.9	1.5	8.5	-2.1	3.
	$D_{2\,wc}$	0.5	3.5	8.5	8.1	1.9	5.2	3.8	5.9	1.8	1.8	7.8	0.9	3.

Table S6. D_{mean} (Gy), D_2 nominal (Gy) and V_{30} (%) scenario selection from marginal probabilities (SSMP). Evaluated conventionally (CE) as seen in section 2.3.

OAR	Unit	Patient												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Esophagus	$D_{2\ nom}$	20.0	61.4	60.4	60.9	61.2	60.5	60.9	61.2	60.7	60.8	60.9	61.5	61.2
	$D_{2\ wc}$	28.3	62.7	63.3	61.9	62.4	61.4	62.8	61.4	61.4	61.0	61.6	61.9	61.5
Heart	D_{mean}	4.9	10.6	6.0	6.8	9.5	3.6	9.8	7.4	1.9	2.5	7.6	4.2	9.8
	V_{30}	7.3	16.0	8.6	9.9	14.6	5.1	14.1	11.2	2.2	3.3	9.1	5.6	15.3
Lungs-GTV	D_{mean}	7.6	22.1	16.0	14.3	14.3	9.0	14.8	11.2	11.5	7.3	12.8	11.8	16.2
	V_{30}	8.3	29.3	25.4	18.9	16.4	12.2	20.1	17.1	17.1	11.3	18.5	17.4	26.5
Spinal cord	$D_{2\ nom}$	16.1	27.2	30.8	34.9	30.0	44.1	24.6	42.2	39.4	18.7	37.7	21.6	30.1
	$D_{2\ wc}$	18.4	32.3	49.1	47.0	30.9	57.3	38.8	56.6	47.2	22.0	54.4	32.8	41.6

Table S7. D_{mean} (Gy), D_2 nominal (Gy) and V_{30} (%) selection from joint probabilities (SSJP). Evaluated conventionally (CE) as seen in section 2.3.

OAR	Unit	Patient												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Esophagus	$D_{2\ nom}$	19.2	60.9	60.6	60.5	60.9	57.4	60.7	60.9	60.5	60.3	60.7	60.7	60.6
	$D_{2\ wc}$	24.7	61.7	61.6	61.3	61.3	60.8	61.8	61.3	61.1	61.1	61.3	61.2	61.2
Heart	D_{mean}	4.3	8.8	4.7	5.3	8.0	2.7	8.6	6.2	1.6	1.9	6.4	3.4	8.6
	V_{30}	6.4	11.6	6.3	7.4	11.9	3.6	12.2	9.1	1.9	2.3	7.3	4.2	13.5
Lungs-GTV	D_{mean}	7.2	20.1	14.7	12.6	12.4	8.0	13.6	10.1	10.8	6.7	12.0	11.0	16.1
	V_{30}	7.6	25.2	23.1	15.8	13.2	10.9	17.4	15.4	15.7	10.4	16.9	16.1	26.4
Spinal cord	$D_{2\ nom}$	15.1	26.9	21.1	27.3	28.1	36.1	22.5	35.3	37.5	17.2	29.2	23.7	26.8
	$D_{2\ wc}$	17.9	28.8	40.6	38.9	29.0	52.1	35.0	50.7	45.4	20.2	46.6	31.9	38.3

References

[1] Chang JY, Zhang X, Knopf A, Li H, Mori S, Dong L, et al. Consensus Guidelines for Implementing Pencil-Beam Scanning Proton Therapy for Thoracic Malignancies on Behalf of the PTCOG Thoracic and Lymphoma Subcommittee. Int J Radiat Oncol Biol Phys 2017;99:41–50. https://doi.org/10.1016/j.ijrobp.2017.05.014.

[2] Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. Phys Med Biol 2012;57. https://doi.org/10.1088/0031-9155/57/11/R99.

[3] Brousmiche S, Souris K, De Xivry JO, Lee JA, Macq B, Seco J. Combined influence of CT random noise and HU-RSP calibration curve nonlinearities on proton range systematic errors. Phys Med Biol 2017;62:8226–45. https://doi.org/10.1088/1361-6560/aa86e9.

[4] Kraus KM, Heath E, Oelfke U. Dosimetric consequences of tumour motion due to respiration for a scanned proton beam. Phys Med Biol 2011;56:6563–81. https://doi.org/10.1088/0031-9155/56/20/003.

[5] Park PC, Cheung JP, Zhu XR, Lee AK, Sahoo N, Tucker SL, et al. Statistical assessment of proton treatment plans under setup and range uncertainties. Int J Radiat Oncol Biol Phys 2013;86:1007–13. https://doi.org/10.1016/j.ijrobp.2013.04.009.

[6] Liu W, Zhang X, Li Y, Mohan R. Robust optimization of intensity modulated proton therapy. Med Phys 2012;39:1079–91. https://doi.org/10.1118/1.3679340.

[7] Pflugfelder D, Wilkens JJ, Oelfke U. Worst case optimization: A method to account for uncertainties in the optimization of intensity modulated proton therapy. Phys Med Biol 2008;53:1689–700. https://doi.org/10.1088/0031-9155/53/6/013.

[8] Fredriksson A, Forsgren A, Hårdemark B. Minimax optimization for handling range and setup uncertainties in proton therapy. Med Phys 2011;38:1672–84. https://doi.org/10.1118/1.3556559.

[9] Bangert M, Hennig P, Oelfke U. Analytical probabilistic modeling for radiation therapy treatment planning. Phys Med Biol 2013;58:5401–19. https://doi.org/10.1088/0031-9155/58/16/5401.

[10] Unkelbach J, Paganetti H. Robust Proton Treatment Planning: Physical and Biological Optimization. Semin Radiat Oncol 2018;28:88–96. https://doi.org/10.1016/j.semradonc.2017.11.005.

[11] Cummings D, Tang S, Ichter W, Wang P, Sturgeon JD, Lee AK, et al. Four-dimensional Plan Optimization for the Treatment of Lung Tumors Using Pencil-beam Scanning Proton Radiotherapy. Cureus 2018;10. https://doi.org/10.7759/cureus.3192.

[12] Inoue T, Widder J, van Dijk L V., Takegawa H, Koizumi M, Takashina M, et al. Limited Impact of Setup and Range Uncertainties, Breathing Motion, and Interplay Effects in Robustly Optimized Intensity Modulated Proton Therapy for Stage III Non-small Cell Lung Cancer. Int J Radiat Oncol Biol Phys 2016;96:661–9. https://doi.org/10.1016/j.ijrobp.2016.06.2454.

[13] Korevaar EW, Habraken SJM, Scandurra D, Kierkels RGJ, Unipan M, Eenink MGC, et al. Practical robustness evaluation in radiotherapy – A photon and proton-proof

alternative to PTV-based plan evaluation. Radiother Oncol 2019;141:267–74. https://doi.org/10.1016/j.radonc.2019.08.005.

[14] Sterpin E, Rivas ST, Van Den Heuvel F, George B, Lee JA, Souris K. Development of robustness evaluation strategies for enabling statistically consistent reporting. Phys Med Biol 2021;66. https://doi.org/10.1088/1361-6560/abd22f.

[15] Buti G, Souris K, Montero AMB, Lee JA, Sterpin E. Towards fast and robust 4D optimization for moving tumors with scanned proton therapy. Med Phys 2019;46:5434–43. https://doi.org/10.1002/mp.13850.

[16] Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures. Med Phys 2016;43:1700–12. https://doi.org/10.1118/1.4943377.

[17] Perkó Z, Voort SR van der, Water S van de, Hartman CMH, Hoogeman M, Lathouwers D. Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion. Phys Med Biol 2016;61:4646. https://doi.org/10.1088/0031-9155/61/12/4646.

[18] Taasti VT, Hattu D, Vaassen F, Canters R, Velders M, Mannens J, et al. Treatment planning and 4D robust evaluation strategy for proton therapy of lung tumors with large motion amplitude. Med Phys 2021;48:4425–37. https://doi.org/10.1002/MP.15067.

[19] Wanet M, Sterpin E, Janssens G, Delor A, Lee JA, Geets X. Validation of the mid-position strategy for lung tumors in helical TomoTherapy. Radiother Oncol 2014;110:529–37. https://doi.org/10.1016/j.radonc.2013.10.025.

[20] Borderías Villarroel E, Geets X, Sterpin E. Online adaptive dose restoration in intensity modulated proton therapy of lung cancer to account for inter-fractional density changes. Phys Imaging Radiat Oncol 2020;15:30–7. https://doi.org/10.1016/j.phro.2020.06.004.

[21] Van Herk M, Remeijer P, Rasch C, Lebesque J V. The probability of correct target dosage: Dose-population histograms for deriving treatment margins in radiotherapy. Int J Radiat Oncol Biol Phys 2000;47:1121–35. https://doi.org/10.1016/S0360-3016(00)00518-6.

[22] Souris K, Barragan Montero A, Janssens G, Di Perri D, Sterpin E, Lee JA. Technical Note: Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy. Med Phys 2019;46:4676–84. https://doi.org/10.1002/mp.13749.

[23] Liu W, Zhang X, Li Y, Mohan R. Robust optimization of intensity modulated proton therapy. Med Phys 2012;39:1079–91. https://doi.org/10.1118/1.3679340.

[24] Liu W, Schild SE, Chang JY, Liao Z, Chang YH, Wen Z, et al. Exploratory Study of 4D versus 3D Robust Optimization in Intensity Modulated Proton Therapy for Lung Cancer. Int J Radiat Oncol Biol Phys 2016;95:523–33. https://doi.org/10.1016/j.ijrobp.2015.11.002.

[25] Witte MG, Sonke JJ, Siebers J, Deasy JO, Van Herk M. Beyond the margin recipe: The

probability of correct target dosage and tumor control in the presence of a dose limiting structure. Phys Med Biol 2017;62:7874–88. https://doi.org/10.1088/1361-6560/aa87fe.