Towards fast and robust 4D optimization for moving tumors with scanned proton therapy.

Gregory Buti[1], Kevin Souris[1], Ana M. Barragán Montero[1], John A. Lee[1], Edmond Sterpin[1,2]

[1]Université Catholique de Louvain, Institut de Recherche Expérimentale et Clinique (IREC), Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Brussels, Belgium

[2]Katholieke Universiteit Leuven, Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium

Version typeset September 2, 2019

E-mail: gregory.buti@uclouvain.be

Abstract

Purpose: Robust optimization is becoming the gold standard for generating robust plans against various kinds of treatment uncertainties. Today, most robust optimization strategies use a pragmatic set of treatment scenarios (so-called uncertainty set) consisting of combinations of maximum errors, of each considered uncertainty source. This approach presents two key issues. First, a subset of considered scenarios are unnecessarily improbable which could potentially compromise the plan quality. Second, the resulting uncertainty set leads to long plan computation times, which limits the potential for robust optimization as a standard clinical tool. In order to address these issues, a method is introduced which is able to pre-select a limited set of relevant treatment error scenarios.

Methods: Uncertainties due to systematic setup errors, image-conversion errors and respiratory organ motion are considered. A 4D-equiprobability hypersurface is defined, which takes into account the joint probabilities of the above-mentioned uncertainty sources. Only scenarios that lie on the pre-defined 4D hypersurface are considered, guaranteeing statistical consistency of the uncertainty set. In this regard, twelve scenarios are selected that cover maximum spatial displacements of the tumor during breathing. Subsequently, additional scenarios are considered (sampled from the aforementioned 4D hypersurface) in order to cover any estimated residual range errors. Two

different scenario-selection procedures were tested: (1) the *maximum displacements* (MD) method that only considers twelve scaled maximum displacement scenarios and (2) *maximum displacements and residual range* (MDR) method which, in addition to the scaled maximum displacement scenarios, considers additional maximum range uncertainty scenarios. The methods were tested for five lung cancer patients by performing comprehensive Monte Carlo robustness evaluations.

Results: A plan computation time gain of 78% is achieved by applying the MD method, whilst obtaining a target robustness of D_{95} larger than 95% of the prescribed dose, for the worst-case scenario. Additionally, MD method has the potential to be fully automatic which makes it a promising candidate for fast automatic planning workflows. The MDR method produced plans with excellent target robustness (D_{99} larger than 95% of the prescribed dose, even for the worst-case scenario), whilst still obtaining a significant plan computation time gain of 57%.

Conclusions: Two scenario-selection procedures were developed which achieved significant reduction of plan computation time and memory consumption, without compromising plan quality or robustness.

Keywords— proton therapy, robust optimization, lung tumors

I. Introduction

Clinical trials have indicated a potential clinical benefit of proton therapy, due to its improved physical dose deposition properties.[1,2,3] Such benefit is related to the steep dose fall-off at the proton's end-of-range (so-called "Bragg peak") which creates the possibility to spare healthy tissues without compromising target coverage. Unfortunately, the high dose gradients make intensity-modulated proton therapy (IMPT) plans sensitive to treatment uncertainties. Important sources of uncertainties include, amongst others, setup errors as well as image-conversion errors (related to the CT image and conversion of the CT Hounsfield units (HUs) to stopping powers). Additionally, tumor motion is another important source of uncertainty which is composed of the following two main elements: (1) changes in the local position of the tumor during delivery (intra-fraction motion), with potential issues related to the interplay effect,[4,5,6] and (2) changes in the average position of the tumor over a respiratory cycle, referred to as a "baseline shift" (with both intra- and inter-fraction components).[7,8] In addition to geometrical uncertainties, the aforementioned errors induce an uncertainty on the estimated proton range, i.e. uncertainty on the position of the Bragg peak, which may cause a deterioration of the actual delivered dose distribution.[9,10,11,12,13,14,15,16] Hence, taking uncertainties into account at the planning stage is critical for successfully treating patients.

To this end, two main robust planning formalisms have been developed: (1) safety margins, and (2) robust optimization. The safety margin approach aims at covering treatment errors by geometrically expanding the "clinical target volume" (CTV) into a "planning target volume" (PTV). A well-known margin recipe is the one developed by van Herk.[17] However, studies have demonstrated that the classic CTV-PTV margin is unable to cover for the range errors in proton therapy; this is due to the failure of the margin recipe's implicitly assumed "static dose cloud approximation" in proton dose distributions.[18,19] Consequently, beam-specific PTVs (BSPTVs) were introduced which adequately account for range uncertainties, under the influence of various treatment errors.[20] Unfortunately, BSPTVs can only be used in single-field uniform dose optimization which is considered inferior to multi-field optimization in proton therapy.[21]

Alternatively, robust optimization methods have been introduced, in which treatment errors are directly incorporated in the optimization process.[22,23,24,25,26] In this study, we focus on a robust optimization method commonly called 'worst-case' robust optimization. Worst-case robust

optimization aims at ensuring adequate target coverage by defining an uncertainty set of treatment error scenarios, defined as the realizations of specific combinations of treatment errors. These error scenarios are evaluated at each iteration of the optimization process with the optimization variables (i.e., the spot weights) adjusted so that the objective function of the current worst-case scenario (the one with the highest value) will be minimized. A popular implementation of worst-case robust optimization is the so-called "minimax" optimization of Fredriksson.[24] Studies demonstrate that worst-case robust optimization can outperform PTV based plans in terms of guaranteeing robustness of the target coverage.[27,28,29]

Two issues are identified in the typical worst-case robust optimization workflow. First, the conventional choice of the uncertainty set limits the ability to handle various types of errors in a statistically sound way. Second, the increased computational burden of the optimization algorithm, related to the high number of required error scenarios, hampers the use of robust optimization in the clinical environment. The availability of computationally cheap algorithms is particularly important in online adaptive workflows, where robust optimization is considered unsuitable due to its long computation time.[30]

More specifically, worst-case robust optimization aims at achieving robustness, by selecting scenarios which represent combinations of maximum errors of each considered uncertainty source, within a pre-defined confidence interval.[24] For instance, a moving lung tumor case typically uses combinations of ±5 mm setup errors in the three directions,[24,31,32] flat image-conversion errors of ±3%[24,15,32] and maximum inhale/exhale breathing phases, giving an uncertainty set of 63 error scenarios (7 setup error scenarios × 3 image-conversion error scenarios × 3 breathing phases). However, this approach is statistically inconsistent as it does not account for the joint probabilities of the considered error sources. Moreover, such approach overlooks the fact that intermediate setup errors could potentially result in even larger range uncertainties.

Additionally, because all error sources are handled in a mutually independent way,[24] an increase of the amount of considered error sources is not practically realizable as this will exponentially increase the size of the uncertainty set. For instance, if baseline shifts or delineation errors are also considered, then the required number of scenarios scale from 63 to hundreds or even thousands scenarios. Attempts have been made to mitigate the need for a large uncertainty set, by deriving empirical formulas which convert robustness parameters of one type of error source into another.[33]

However, this solution is limited as evaluations for a different tumor location requires re-evaluation of the recipe.

This study aims at establishing a scenario-selection procedure that addresses the above-mentioned issues. The focus lies in an efficient pre-selection of a limited number of relevant error scenarios, which are later on fed to a worst-case robust optimizer. As will be illustrated, the resulting uncertainty set contains scenarios that are statistically consistent, whilst its reduced size limits the computational burden of the optimization process.

II. Material and Methods

In this section, first the statistical framework is presented, followed by a detailed explanation of the proposed methods and reference method. Afterwards, we give an overview of the planning and evaluation software applied for testing the respective methods. Finally, the section concludes with a description of the patient data and the quality metrics for the evaluation and comparison of the treatment plans.

II.A. Methodology

II.A.1. Statistical Framework

Uncertainties due to systematic setup errors, image-conversion errors and respiratory organ motion are considered. Because the organ motion is represented by a set of equally spaced phases in time (see Section II.E.), each phase is assumed to be equally probable.

The systematic setup errors $\boldsymbol{x_s} = (x_s, y_s, z_s)$ along left-right x, anterior-posterior y and superior-inferior z directions are assumed to be described by a 3D-Gaussian probability distribution (characterized by a standard deviation $\boldsymbol{\Sigma_s} = (\Sigma_{xs}, \Sigma_{ys}, \Sigma_{zs})$).[a] By following Van Herk's margin recipe,[17] a confidence interval for the above-mentioned 3D distribution is generated by considering all setup errors that satisfy the following inequality:

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 \leq \alpha_{3D}^2, \tag{1}$$

with α_{3D} being a coverage parameter that can be adapted to specify the integration limit in the error scenario space, or in other words, to fix the width of the confidence interval. Values for α

[a] Bold symbols represent vectors.

in 1D, 2D and 3D can be found in Van Herk.[17] For the general N-dimensional case, the following formula can be used to evaluate α_{ND} numerically:[34]

$$\alpha_{ND} = \sqrt{\text{inv-}\chi^2(C, N)}, \tag{2}$$

with C the confidence interval and inv-χ^2 the inverse cumulative density function of the chi-squared distribution. Equation 2 was evaluated with Matlab in order to obtain the different values for α_{ND}. For a perfect 3D dose conformation of the target, the clinically recommended confidence interval is 90%, which corresponds to a value for α_{3D} of 2.5. A 3D-equiprobability hypersurface can subsequently be constructed by regarding the maximum setup errors, limited by the inequality in Equation 1:

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 = \alpha_{3D}^2. \tag{3}$$

In proton therapy planning, image-conversion errors must also be handled. In contrast to setup errors, image-conversion errors $\boldsymbol{r}$ only vary in one dimension and are thus described by a 1D-Gaussian probability distribution (characterized by sigma $\boldsymbol{\Sigma_r}$).[9] Hence, if both setup errors and image-conversion errors are considered, the probability of a treatment error scenario (defined as a specific combination of a setup error and image-conversion error) has to be treated with increased dimensionality as compared to the confidence interval that defines the hypersurface of Equation 3. As a result, the probability distribution that describes the treatment error realizations is four-dimensional and the scenarios that lie within the pre-defined confidence interval (in scenario space), are represented by:

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 + \left(\frac{r}{\Sigma_r}\right)^2 \leq \alpha_{4D}^2. \tag{4}$$

In this case, the 90% confidence interval is represented by a value for α_{4D} of 2.8 (using Equation 2). The inequality of Equation 4 defines the following 4D-equiprobability hypersurface:

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 + \left(\frac{r}{\Sigma_r}\right)^2 = \alpha_{4D}^2. \tag{5}$$

Hence, we can sample equiprobable scenarios (x_s, y_s, z_s, r), i.e. specific combinations of setup errors and image-conversion errors, which are positioned exactly on the edge of the pre-defined confidence interval. Two conditions are defined which must be satisfied by the considered scenarios:

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 \leq \alpha_{3D}^2, \tag{6}$$

$$\left(\frac{x_s}{\Sigma_{xs}}\right)^2 + \left(\frac{y_s}{\Sigma_{ys}}\right)^2 + \left(\frac{z_s}{\Sigma_{zs}}\right)^2 + \left(\frac{r}{\Sigma_r}\right)^2 = \alpha_{4D}^2. \tag{7}$$

The first condition (Equation 6) restricts the magnitude of the setup errors and is identical to the condition that yields the margin recipe (Equation 3). Hence, the spatial displacements of the CTV will be limited by the maximum considered setup error. The second condition (Equation 7) guarantees that only scenarios of equal probability, defined by the coverage parameter α_{4D}, are selected. The 90% equiprobability line, from which the scenarios are sampled, is shown in Figure 1. As illustrated in the figure, the constraint of the maximum setup error, imposed by the inequality of Equation 6, reduces the considered confidence interval in scenario space. A maximum setup error of 5 mm is chosen in order to limit the maximum setup error to a value commonly found in other worst-case robust optimization studies, see for example[24,31,22,32]. Nevertheless, we must rely on an unbiased robustness evaluation to check if the treatment plan satisfies the robustness criteria as defined by the confidence interval in dosimetric space.

Values for the setup error standard deviation $\Sigma_{xs} = \Sigma_{ys} = \Sigma_{zs}$ are set equal to 2 mm in order to provide a uniform maximum setup error of 5 mm ($= x_{s,max} = y_{s,max} = z_{s,max}$), at a 90% confidence interval[b]. Following the review of Paganetti,[9] the magnitude of the image-conversion error standard deviation Σ_r is set equal to 1.6% (this value was reported for calculations with a Monte Carlo dose engine).

II.A.2. Scenario-Selection Procedures

Using the formulation described in Section II.A.1., two different procedures of selecting relevant error scenarios are investigated: (1) *maximum displacements* method (MD) and (2) *maximum displacements and residual range* method (MDR). Both procedures are described in detail below. Afterwards, the performance of the two proposed scenario-selection methods (MD and MDR) will be compared to the conventional robust optimization (without pre-selection of scenarios), where the treatment plans are constructed using an uncertainty set of 63 scenarios, i.e. combinations of ±5 mm setup errors in the three directions, ±2.6% image-conversion error (see Section II.B.) and maximum inhale/exhale breathing phases (as it would be performed conventionally in commercial TPSs).

[b] using Equation 6, $\Sigma_{xs} = x_{s,max}/\alpha_{3D}$, with $x_{s,max} = 5$ mm and $\alpha_{3D} = 2.5$ at a 90% confidence interval (analogous for the other directions $y_{s,max}$ and $z_{s,max}$).

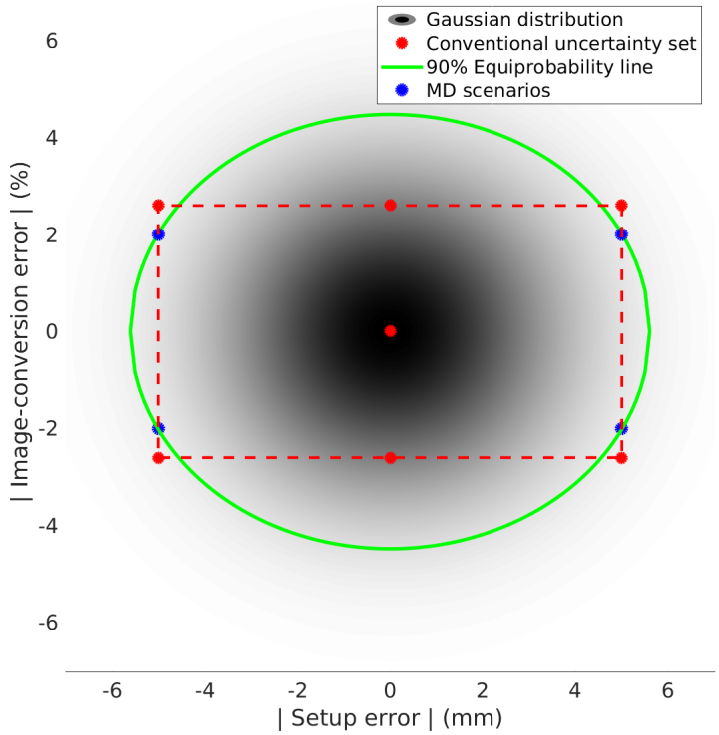


Figure 1: 2D-Gaussian probability distribution, defined by $\|\boldsymbol{\Sigma_s}\|$ and $\|\Sigma_r\|$, representing the likelihood of sampled scenarios (the lighter, the more unlikely). The 90% equiprobability line (green) defines all possible scenarios that are positioned exactly on the edge of the 90% confidence interval. The scenarios within the conventional uncertainty set (combinations of ± 5 mm setup errors and flat $\pm 3\%$ image-conversion errors) are depicted by the red circles (7 scenarios in 2D). The maximum displacement (MD) scenarios (as explained in Section II.A.2.) are depicted by the blue circles (4 scenarios in 2D).

Maximum displacements (MD)

In the MD method, twelve scenarios are selected that aim to cover the extreme positions reached by the tumor. If respiratory motion is considered, these scenarios are determined as follows: first, the target centers of mass are computed for all breathing phases. Then, six phases are selected where the center of mass reaches its maximum value, along along the three directions ($\pm x$, $\pm y$ and $\pm z$). For each of the resulting six phases, a maximum setup error ($= 5$ mm), in the direction of largest spatial displacement is applied, by rigidly shifting the chosen CT images. For example, in the breathing phase with largest displacement in the $+x$ direction, a setup error of $+x_s = (+5\text{mm},0,0)$ is applied. Analogously for the other directions. In the case of non-moving tumors, the six maximum displacement scenarios are simply represented by the maximum setup error along $\pm x$, $\pm y$ and $\pm z$

directions. Finally, to each scenario, an image-conversion error is applied with a magnitude equal to the maximum value $\pm r$ allowed by the 4D-equiprobability hypersurface (Equation 7). That is, each of the six scenarios are scaled with both positive and negative image-conversion errors $\pm r$ (equal to $\pm 2\%$), providing twelve scenarios in total.

The application of image-conversion errors on the CT image is performed by uniformly scaling the mass densities obtained from the CT image (using the same CT calibration curve as in the dose calculation). The twelve scaled maximum spatial displacement scenarios can be interpreted by the intersection of the 90% equiprobability line with the box, which is constructed by the scenarios of the conventional uncertainty set, at the 5 mm setup error. The uncertainty set of the MD method, contains thirteen scenarios (twelve selected scenarios in addition to the nominal scenario (= planning CT)). Each selected error scenario is simulated by modifying the original CT with the chosen error values, generating virtual CTs that will later be imported in the treatment planning system (TPS).

Maximum displacements and residual range (MDR)

In the MDR method, in addition to the MD scenarios, additional scenarios are considered which have estimated range errors larger than the ones induced by the twelve MD scenarios already present in the uncertainty set. In other words, we want to include scenarios that will cover any residual range errors, i.e. range errors that are not yet covered by previously included scenarios. These scenarios are selected as follows: first, proton ranges can be estimated by converting the considered breathing CT images into maps of water-equivalent path lengths (WEPLs).[c] Because WEPLs are beam-specific, each breathing phase has a separate WEPL map for each respective beam angle. Scenarios are then simulated by sampling treatment errors as follows:

- Random selection of a breathing phase and beam angle, as well as,
- Random sampling of a combination of setup error (x_s,y_s,z_s) and image-conversion error r that satisfies both Equations 6 and 7.

[c]The WEPL in a voxel is obtained by integrating the relative stopping power ratio (RLSP) of the voxels along the beam path: $WEPL = \int_0^L RLSP(HU, l)dl$ for each beam angle. WEPL maps are computed using the open-source platform OpenReggui[35] which uses a fast ray-tracing algorithm[36] for its WEPL calculations.

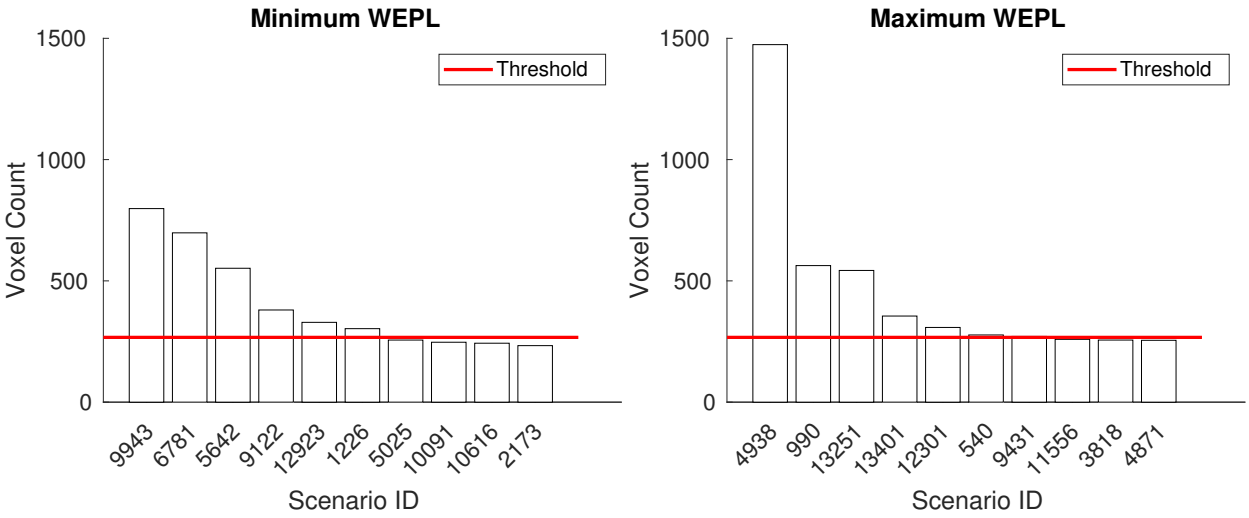
The sampling of breathing phases can be omitted if breathing motion is not considered. For each scenario, the sampled setup error is applied by rigidly translating the pre-computed WEPL map image. For the image-conversion error, the WEPL values are scaled with the respective error value r. By repeating this process, a distribution of WEPL values for all target voxels is obtained across all scenarios. Finally, a voxel-based scenario selection is performed by identifying which scenario s has induced the largest residual range for most of the target voxels (see Figure 2). To compute this, the following four matrices are stored. First, the maximum and minimum WEPLs, for each target voxel, across the MD scenarios, are stored in W_{MD}^{max} and W_{MD}^{min}, respectively. Second, the maximum and minimum WEPLs, for each target voxel, across all randomly sampled scenarios, are stored in W_{rand}^{max} and W_{rand}^{min}, respectively. Afterwards, we can identify worst-case overshoot scenarios by computing for each randomly sampled scenario, the number of voxels N_{max} that it has in common with W_{rand}^{max} *and* that induce WEPL values larger than W_{MD}^{max}. Analogously, worst-case undershoot scenario are classified according to the number of voxels N_{min} that each sampled scenario has in common with W_{rand}^{min} *and* smaller than W_{MD}^{min}:

$$N_{max} = \#\{n_i \mid W_s(\boldsymbol{i}) = W_{rand}^{max}(\boldsymbol{i}) \;\&\; W_s(\boldsymbol{i}) > W_{MD}^{max}(\boldsymbol{i})\}_{\boldsymbol{i}\in \mathrm{CTV}}, \tag{8}$$

$$N_{max} = \#\{n_i \mid W_s(\boldsymbol{i}) = W_{rand}^{min}(\boldsymbol{i}) \;\&\; W_s(\boldsymbol{i}) < W_{MD}^{min}(\boldsymbol{i})\}_{\boldsymbol{i}\in \mathrm{CTV}}, \tag{9}$$

with n_i an auxiliary variable, W_s the WEPL map of scenario s and $\boldsymbol{i}$ the vector that represents the voxels in the CTV. In other words, worst-case scenarios are selected in which the combination of setup errors, image-conversion errors and breathing phases have estimated proton ranges that deviate most from the values in the previously included scaled maximum spatial displacement scenarios.

In order to limit the size of the uncertainty set, we define a threshold (Figure 2) that discards scenarios which induce maximum residual ranges in less than 2% of target voxels ($= 2\%N_{CTV}$ with N_{CTV} the total number of CTV voxels). Using Equations 8 and 9, the scenarios that do not meet $N_{max} < 2\%N_{CTV}$ and $N_{min} < 2\%N_{CTV}$ are discarded for the overshoot and undershoot scenarios, respectively. By doing so, we avoid the selection of scenarios that cover only few range errors (see Discussion in Section IV.). As a result, the MDR method's uncertainty set contains the twelve maximum displacement scenarios, with additional error scenarios that aim at covering any residual range errors. Analogous to the MD method, virtual CTs are generated that represent the selected error scenarios.



(a) Undershoot scenarios (b) Overshoot scenarios

Figure 2: Illustration of the voxel based scenario selection. For example, scenario with ID s induces a worst-case range error for y number of voxels in the target volume (maximum WEPL for overshoot and minimum WEPL for undershoot). Hence, scenarios are ordered according to the maximum range error they induced in most target voxels (left panel: worst-case undershoot scenarios, right panel: worst-case overshoot scenarios).

It must be noted that, in the scenario-selection procedure, the calculation of the WEPL maps consumes the largest share of the total pre-computation time. Moving lung tumor cases, together with three beam plans, require 69 WEPL maps (11 breathing phases + 12 MD scenarios, each with three beam angles). For a single scenario, the calculation of a WEPL map takes approximately 6 seconds for smaller target volumes ($\sim$41 cm^3) and 15 seconds for a deep-seated larger target volume ($\sim$152 cm^3), amounting to an upper limit of 17 minutes. Moreover, once the WEPL maps are stored, errors scenarios are generated quasi instantaneously. The advantage of this approach is that it does not involve any dose evaluations and, hence, many scenarios ($>10^4$) can be evaluated in a very short time period. Sampling and evaluation of 10^4 scenarios typically takes less than 2 minutes. Together with the WEPL map calculations and scenario creation (max. 4 minutes), this gives a maximum pre-computation time of 23 minutes.

II.B. Treatment Planning System

Treatment plan optimization is performed with the 4D-robust optimization algorithm of the TPS RayStation research version v7.99 (RaySearch Laboratories, Stockholm, Sweden). The time-averaged mid-position CT is used as the nominal planning CT which was created with the open-

source platform OpenReggui.[37,35] OpenReggui calculates the mid-position CT by computing the mean position over the respiratory cycle after deformable registration between all phases of the 4D-CT image set. The Monte Carlo dose engine of the TPS is used for the dose calculations with 10^4 ions per spot and a 3×3×3 mm^3 dose calculation grid.

For the *conventional* method, the robust optimization tool of the TPS is used, selecting robustness parameters of 5 mm setup errors in all directions, 2.6% image-conversion errors and maximum inhale and maximum exhale phases (total of 63 scenarios). A value of 2.6% is chosen because it represents the value at which 90% of image-conversion errors are covered, assuming they are described by a 1-D Gaussian distribution, i.e. $2.6\% = \alpha_{1D}\Sigma_r$ with $\alpha_{1D} = 1.64$ (Equation 2) and $\Sigma_r = 1.6\%$. As mentioned in Section II.A., treatment plans of the MD and MDR methods are obtained by importing the DICOM CT data of the virtual CTs in the TPS, which represent the selected set of error scenarios. A 4D-robust plan optimization is then performed over the imported CT images.

II.C. Evaluation Software

Treatment plans are evaluated with the independent Monte Carlo dose engine MCsquare, available open-source.[38] MCsquare has been commissioned and validated for clinical practice. The same beam model (optimised from the commissioning measurements) was used for the Monte Carlo and TPS dose calculations, thus avoiding possible errors due to algorithm-machine calibration. The dose level difference (evaluated at D_{95}) between a MCsquare and the TPS is typically less than 0.1 Gy, for final dose calculation at a 1% statistical uncertainty.

The effects of systematic setup errors, image-conversion errors and breathing motion on the planned dose distribution are evaluated by performing comprehensive robustness evaluations with MCsquare.[39] In each robustness test, a set of 250 error scenarios were sampled with the number of protons selected in order to reach a statistical uncertainty of 1%.

MCsquare follows a Monte Carlo approach for its robustness evaluation, by randomly sampling error scenarios according to the error distributions mentioned below.[40] For all error scenarios, the dose distributions are recomputed, discarding the 10% worst scenarios (based on the target D_{95}). Because scenarios are sampled from the entire dosimetric error space, the selection of evaluation scenarios is not limited by the 90% equiprobability hypervolume in the scenario space, utilized for

the selection of the optimization scenarios (see Section II.A.1.). Hence, the robustness tests can be considered as an unbiased representation of the plan's sensitivity to the treatment errors.

Probability distributions for setup errors and image-conversion errors are identical to the distributions used in the planning process (standard deviations of 2 mm and 1.6% for setup and image-conversion errors, respectively). MCsquare models the setup errors and image-conversion errors by rigidly translating the CT image (= shifting the beam isocenter) for the first one, whilst scaling the CT densities for the latter. Breathing motion is simulated by recomputing the dose distribution for each breathing phase and accumulating the dose on the mid-position CT.

II.D. Patient Cases

Lung tumor cases were chosen with the purpose of testing the proposed methods, as they typically present difficulties in terms of ensuring target robustness (large density heterogeneities and large tumor motion). Treatment plans were calculated for five lung tumor patients, all diagnosed with single tumor volume, delineated on the CT data. The set of patients presented a wide range of varying tumor size and motion amplitude, therefore representative of the entire patient population. Patient data were characterized by a 4D-CT image set, binned in ten breathing phases, equally spaced in time. The main features of the patient cohort are summarized in Table 1. All treatment plans were designed using a configuration of three co-planar fields, delivered via IMPT with the pencil beam scanning (PBS) technique (see Table 1).

Treatment plans were constructed with identical target and OARs objectives in the optimization. Patients had a dose prescription of 60 Gy to the CTV. Target coverage was considered acceptable if 95% of the CTV received more that 95% of the prescribed dose (D_{presc}), whilst no more than 5% of the CTV received over 105% of D_{presc}, even for the worst-case scenario. However, in order to test the proposed methods, we focus on target coverage during the optimization, by aiming to reach CTV $D_{99} \geq 95\% D_{presc}$, in the nominal case.

Table 1: Patient characteristics including tumor size, tumor motion amplitude (in left-right (LR), anterior-posterior (AP) and superior-inferior (SI) directions), tumor position (right-middle lobe (RML), left-lower lobe (LLL), right-upper lobe (RUL), left-upper lobe (LUL)) and beam configuration.

Patient	**CTV size**	**Motion Amplitude**			**Tumor position**	**Gantry angles**
		LR	AP	SI		
	[cm^3]	[mm]	[mm]	[mm]		[°]
P1	152.6	4.2	2.1	3.1	RML	0, 270, 310
P2	107.7	3.1	2.9	3.7	LLL	90,135, 180
P3	41.3	1.4	2.9	0.8	RUL	180, 225, 270
P4	70.3	0.8	1.2	0.5	LUL	90, 135, 180
P5	109.6	2.2	1.8	6.6	RUL	180, 225, 270

III. Results

By comparing target coverage and OAR dose, the methods are assessed for their quality and robustness and their ability to spare the normal tissues. The coverage metrics for the relevant regions-of-interest (ROIs), are derived from the DVHs of the plan's robustness evaluation. The results of the nominal plans were normalized by applying a correction factor in such a way that 50% of the target volume received the prescribed dose. The evaluation dose distributions, for each patient, were scaled with its respective correction factor. The lung, bronchus and heart received significant dose levels and are therefore the OARs reported in the figures and tables.

Figure 3 illustrates the result of the robustness test by displaying the DVH bands of the CTV, lung, bronchus and heart along with the nominal DVHs, for a single patient. The results for the other patients are presented in Tables 4 and 5. The results are concentrated in a summary table (Table 2), displaying for each metric the difference between the value obtained by the *conventional* method with MD method, averaged across all patients and analogously, the difference between the *conventional* method and MDR method. For each evaluation metric, the results are reported in respectively, the average, worst-case and nominal scenarios.

In terms of target coverage, results show treatment plans obtained from all methods passed the target coverage acceptability limit of worst-case $D_{95} \geq 95\%D_{presc}$. Only the MDR and *conventional* methods exceeded a target coverage of $D_{99} \geq 95\%D_{presc}$, in the worst-case scenario, for all patients. Comparing the MDR method with the *conventional* method shows that a similar target coverage is

obtained (average reduction of only 0.1 Gy D_{99} for the worst-case scenario) whilst improving slightly the normal-tissue sparing (sparing of the lung, on average, 1.9% and 0.9 Gy for V_{20} and D_{mean}, respectively and, on average, reducing maximum bronchus dose 0.3 Gy, evaluated for the worst-case scenario). In order to evaluate the plan's sensitivity to the treatment errors, the dose homogeneity of the target volume is calculated by subtracting the worst-case CTV D_{98} from the worst-case CTV D_2 (see Table 3). In general, MDR method produced plans closest to the *conventional* method in terms of homogeneity (an average difference of only 0.2 Gy between both methods).

Table 3 reports the plan computation times, together with the simulated number of scenarios. Results show that the MD method achieved an average time gain of 78% with respect to the *conventional* method. By using the MDR method, the number of optimization scenarios is reduced by approximately a factor of three, on average, which translated in an average time gain of 57%.

Table 2: Difference of the average (across all patients) target and organ-at-risk DVH metrics between plans of the MD with the *conventional* method (MD-Ref) and difference of the average metrics between the MDR with the *conventional* method (MDR-Ref).

	CTV					
	ΔD_{99} [Gy]		ΔD_{95} [Gy]		ΔD_{5} [Gy]	
	MD-Ref	MDR-Ref	MD-Ref	MDR-Ref	MD-Ref	MDR-Ref
Avg.	-0.5	-0.1	-0.1	0.0	0.0	0.0
Worst	-1.6	-0.1	-0.5	0.0	0.0	0.0
Nom.	-0.2	0.0	-0.1	0.0	0.0	0.0

	Lung				**Bronchus**		**Heart**	
	ΔV_{20} [%]		ΔD_{mean} [Gy]		ΔD_{max} [Gy]		ΔV_{40} [%]	
	MD-Ref	MDR-Ref	MD-Ref	MDR-Ref	MD-Ref	MDR-Ref	MD-Ref	MDR-Ref
Avg.	-2.2	-1.7	-1.4	-0.9	-0.4	-0.1	-0.4	-0.2
Worst	-2.8	-1.9	-1.4	-0.9	-0.7	-0.3	-0.4	-0.2
Nom.	-2.7	-1.7	-1.3	-0.8	-0.3	0.0	-0.4	-0.2

Table 3: Plan computation time, number of scenarios and dose homogeneity for plans of each patient (P), obtained using the *conventional* (Ref), MD and MDR methods. The average time differences Δt and average dose homogeneity, across all patients, are reported at the bottom. For the reference method, the plan computation time comprises only of the plan calculation time (= mainly dose-influence matrix calculations and plan optimization). For the MD and MDR method, the total computation time is reported as the pre-computation time + the plan calculation time. The pre-computation time consists of the scenario creation (both MD and MDR methods), WEPL map calculation and scenario sampling (only MDR method).

	Computation time [min]			**Scenarios**			**Dose Homogeneity** [Gy]		
	Ref	MD	MDR	Ref	MD	MDR	Ref	MD	MDR
P1	*229*	2 + 41 = 43	22 + 73 = 95	*63*	13	21	*2.8*	5.1	3.5
P2	*156*	2 + 32 = 34	21 + 44 = 65	*63*	13	15	*2.1*	3.6	2.7
P3	*58*	2 + 12 = 14	10 + 21 = 31	*63*	13	20	*3.0*	3.8	3.0
P4	*94*	2 + 19 = 21	13 + 32 = 45	*63*	13	22	*2.5*	3.7	2.3
P5	*141*	2 + 28 = 30	21 + 33 = 54	*63*	13	15	*2.8*	2.8	2.6
Avg.		Δt = -78%	Δt = -57%	*63*	13	19	*2.6*	3.8	2.8

IV. Discussion

The rationale for introducing a scenario-selection procedure was twofold:

First, the scenario-selection procedure guarantees statistical consistency across scenarios present in the uncertainty set. As Fig. 1 illustrates, the conventional uncertainty set (resulting from the use of a flat $\pm$2.6% image-conversion error) contains scenarios that are positioned outside the equiprobability line. The proposed methods (MD and MDR) do not emphasize these unlikely scenarios and only select equiprobable scenarios that lie within the pre-defined confidence interval, which is set at 90%.

Second, the scenario-selection procedure allows for a reduction of the size of the uncertainty set. Reducing the uncertainty set is important as, for a given patient, the number of input optimization scenarios is directly proportional to the plan computation time (see Figure 4). The main reason for this is that the amount of beamlet dose-influence matrices must be computed and stored for each optimization scenario. Moreover, fewer dose evaluations, at each iteration, improve the speed of the optimization process and reduce the memory consumption. Deciding the optimal robust planning

method will depend on the intended goals of the planning workflow: (A) fast and automatic planning, or (B) robust target coverage.

(A) If focus lies on limiting the computation time, then the time-gain can be maximised by applying the MD method, provided that a target robustness of $D_{95} \geq 95\%D_{presc}$ is deemed acceptable. An additional benefit of this method is its potential to be fully automatic and the fact that the number of pre-computations are limited. In its current implementation, selected error scenarios must be imported manually. However, this can easily be implemented in most commercial TPSs which provide standard scripting tools.

(B) If focus lies on target coverage, then the robustness of the treatment plan can be increased by utilizing the MDR method. Results show that target robustness is significantly improved ($D_{99} \geq 95\%D_{presc}$) whilst still achieving a time gain of 57%, on average. These results indicate that by considering an additional number of estimated worst-case error scenarios, robustness criteria can be satisfied whilst avoiding overly robust solutions. The two main disadvantages of the MDR method are: (1) the necessity of a pre-computation process outside of the TPS (mainly WEPL map calculations), and (2) a prior analysis in order to fix the value of the coverage threshold (see Section II.A.2.). Retrospective analysis found that (see Figure 4), based on the population of patients in this study, discarding scenarios that do not induce residual ranges for more than 2% of target voxels ($N_{max} < 2\%N_{CTV}$ and $N_{min} < 2\%N_{CTV}$, see Equations 8 and 9) resulted in an optimal balance between the number of selected scenarios and the amount of covered range errors. As Figure 4 shows, a more conservative approach may be employed by reducing this threshold even further, with a corresponding increase in the number of selected scenarios. However, because WEPL map evaluations treat each beam angle separately, the effect of the treatment errors in the WEPL space can be considered more substantial than its corresponding effect in the real dosimetric space. Hence, this threshold is deemed satisfactory in order to achieve the necessary robustness of the treatment plan.

The present study focused on moving lung tumor cases where the aim was to achieve robustness against systematic setup errors, image-conversion errors and breathing motion. Random errors should also be considered as they present an important source of range uncertainties. However, random errors require the simulation of fractionation effects for which a pre-selection of optimization scenarios does not suffice. Solutions dealing with random errors simulate their effect during the plan calculation. However, because access to the source code of the TPS is restricted, random

errors have been omitted from the evaluation. In the literature, the following solutions exist which could potentially be used in conjunction with the scenario-selection methods: (1) random errors can be simulated in the Monte Carlo calculations of the beamlet dose-influence matrices, under the assumption of an infinite number of fractions,[41] and (2) the method by Fredriksson[42] can be employed which modifies the optimization objective function in order to include random errors, for a finite number of fractions.

The scenario-selection procedure provides a method for handling other yet unconsidered systematic error sources, within a statistically consistent framework. However, these potential error sources, such as baseline shifts or anatomical changes, should be able to be realistically modeled by creating virtual CTs (analogous to setup and range errors). Furthermore, the method does not change the fundamental worst-case robust optimization algorithm. It can therefore be integrated in any robust planning workflow where a TPS is used that is able to perform 4D-robust optimization.

V. Conclusions

This study introduces a scenario-selection procedure which enables the reduction of the uncertainty set used in worst-case robust optimization. Relevant optimization scenarios are selected according to: (1) maximum spatial displacements of the tumor, and (2) largest estimated range uncertainties. Based on the scenario-selection procedure, two pre-selection methods are proposed and tested for moving lung tumor cases as follows:.

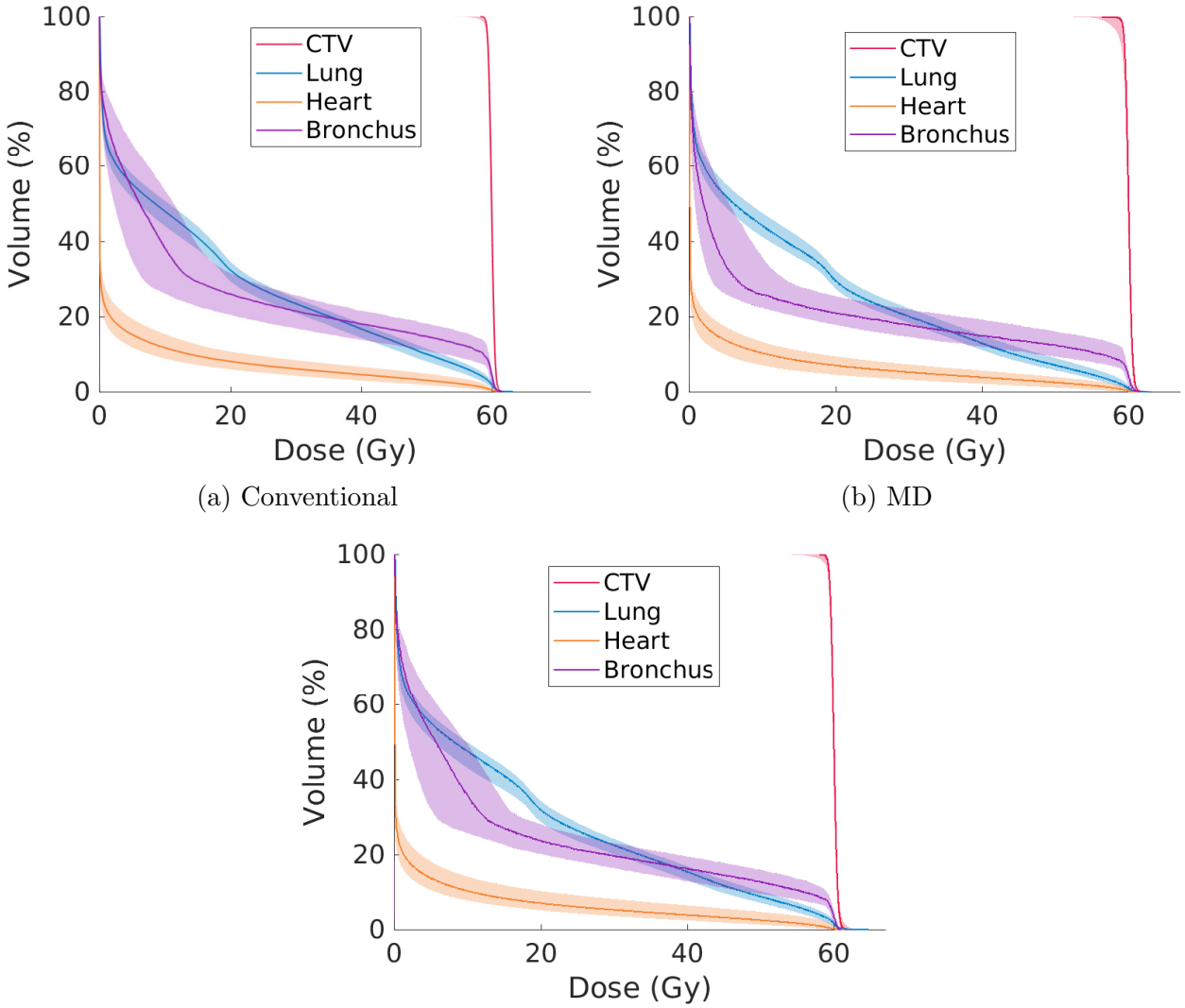
First, the *maximum spatial displacements* (MD) method only considers scenarios corresponding to the maximum spatial displacements of the tumor during breathing, with CT-HU values scaled according to the image-conversion error defined by a pre-defined 4D-equiprobability hypersurface. Because its uncertainty set contains thirteen scenarios (twelve selected scenarios together with the nominal scenario), a reduction of 78% plan computation time is achieved. Moreover, the MD method has the potential to be fully automatic which makes it a promising candidate for fast automatic planning workflows. Second, the *maximum displacements and residual range* (MDR) method is proposed, which adds additional scenarios to the uncertainty set in order to cover for any residual range errors. Results show that this method produces plans with target robustness of CTV $D_{99} \geq 95\%D_{presc}$, whilst achieving a 57% reduction of plan computation time with respect to the sixty-three scenario *conventional* method. Future efforts will concentrate on extending the

scenario-selection procedure by including additional uncertainty sources. This will provide useful insights on the full robust picture and is topic of future research.

Acknowledgements

Gregory Buti is supported by the Télévie Grant from the Belgian 'Fonds National pour la Recherche Scientifique' F.R.S-FNRS (Grant No. 7453918F). Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the F.R.S.-FNRS under convention 2.5020.11. Kevin Souris is supported by a research grant from Ion Beam Application (IBA s.a., Louvain-la-Neuve, Belgium). John A. Lee is a Research Associate with the F.R.S.-FNRS.

Additional Figures and Tables



(a) Conventional

(b) MD

(c) MDR

Figure 3: DVH bands for the CTV, lung and bronchus for plans obtained using the (a) *conventional*, (b) *maximum displacements* (MD) and (c) *maximum displacements and residual range* (MDR) methods, for a single patient (Patient 2).

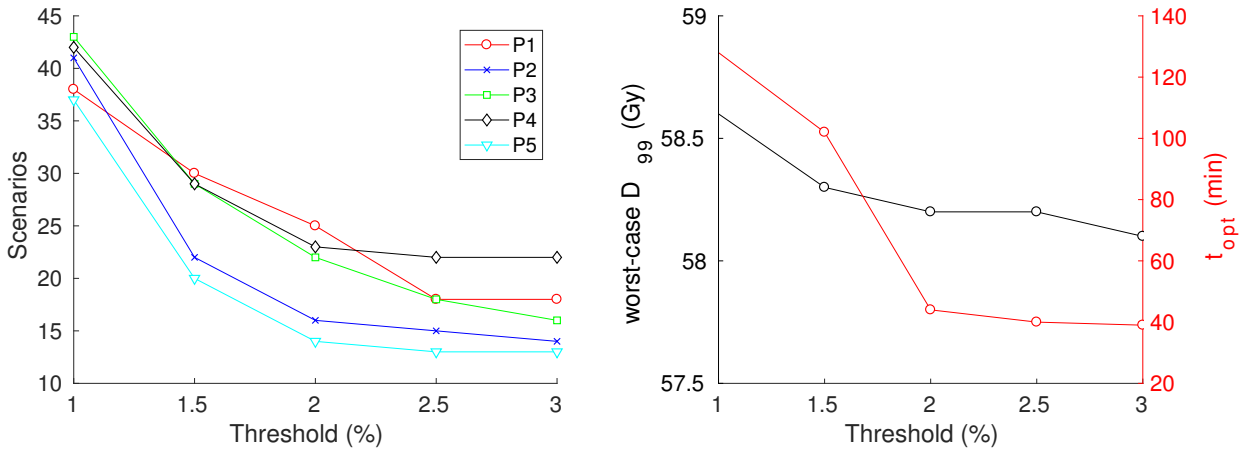


Figure 4: Effect of the a certain threshold value. Left: influence on the number of selected scenarios in the MDR method for patients 1 to 5. Right: example of the influence on the resulting treatment plan (worst-case D_{99} and plan optimization time t_{opt}).

Table 4: Target DVH metrics for plans of each patient (P), obtained using the *conventional* (Ref), *maximum displacements* (MD) and (c) *maximum displacements and residual range* (MDR) methods.

		CTV								
		D_{99} [Gy]			D_{95} [Gy]			D_5 [Gy]		
		Ref	MD	MDR	Ref	MD	MDR	Ref	MD	MDR
P1	Avg.	*58.8*	57.4	58.4	*59.2*	58.9	59.1	*60.8*	60.7	60.8
	Worst	*57.6*	54.7	57.0	*59.0*	57.9	58.8	*60.9*	60.9	61.0
	Nom.	*58.9*	58.1	58.8	*59.3*	59.1	59.2	*60.8*	60.7	60.8
P2	Avg.	*59.0*	58.5	58.9	*59.3*	59.2	59.3	*60.7*	60.7	60.7
	Worst	*58.4*	56.6	*58.2*	*59.2*	58.6	59.2	*60.8*	60.8	*60.8*
	Nom.	*59.1*	59.0	59.1	*59.3*	59.3	*59.4*	*60.7*	60.7	*60.6*
P3	Avg.	*58.5*	58.1	58.5	*59.0*	58.8	59.0	*60.9*	60.8	60.8
	Worst	*57.6*	56.6	57.5	*58.8*	58.4	58.8	*61.0*	61.0	60.9
	Nom.	*58.6*	58.5	58.7	*59.1*	59.0	59.1	*60.8*	60.8	60.7
P4	Avg.	*58.8*	58.7	58.9	*59.2*	59.1	59.2	*60.7*	60.7	60.7
	Worst	*57.5*	56.2	58.3	*59.1*	58.7	59.1	*60.8*	60.8	60.8
	Nom.	*59.0*	59.0	59.0	*59.3*	59.3	59.3	*60.7*	60.8	60.8
P5	Avg.	*58.8*	58.7	58.7	*59.2*	59.2	59.2	*60.8*	60.8	60.7
	Worst	*58.2*	57.2	58.0	*59.1*	59.0	59.0	*60.9*	60.8	60.8
	Nom.	*58.8*	58.9	58.9	*59.2*	59.2	59.3	*60.8*	60.7	60.7

Table 5: Organ-at-risk DVH metrics (lung, bronchus and heart) for plans of each patient (P), obtained using the *conventional* (Ref), *maximum displacements* (MD) and (c) *maximum displacements and residual range* (MDR) methods.

		Lung						**Bronchus**			**Heart**		
		V_{20} [%]			D_{mean} [Gy]			D_{max} [Gy]			V_{40} [%]		
		Ref	MD	MDR	Ref	MD	MDR	Ref	MD	MDR	Ref	MD	MDR
P1	Avg.	*36.3*	30.8	32.1	*16.8*	14.7	15.3	*62.8*	62.5	63.3	*2.8*	1.9	2.6
	Worst	*39.1*	33.6	34.8	*18.1*	15.8	16.3	*63.4*	63.1	63.9	*3.6*	2.7	3.4
	Nom.	*36.7*	31.1	32.6	*17.0*	14.8	15.6	*62.6*	62.7	63.5	*2.9*	2.0	2.7
P2	Avg.	*32.1*	29.2	31.3	*16.4*	14.5	15.8	*61.2*	61.2	61.2	*4.3*	3.7	3.8
	Worst	*33.9*	30.9	32.7	*17.2*	15.4	16.5	*61.5*	61.6	61.6	*5.6*	4.9	5
	Nom.	*32.3*	29.4	31.4	*16.5*	14.6	15.9	*61.7*	61.5	61.0	*4.5*	3.9	3.9
P3	Avg.	*14.3*	13.5	14.0	*7.8*	7.1	7.4	*61.1*	61.0	60.9	*0.0*	0.0	0.0
	Worst	*15.1*	14.4	14.9	*8.2*	7.7	7.9	*61.5*	61.5	61.4	*0.0*	0.0	0.0
	Nom.	*14.4*	13.6	14.1	*7.8*	7.2	7.5	*61.3*	60.9	60.9	*0.0*	0.0	0.0
P4	Avg.	*21.5*	20.3	21.0	*11.2*	10.5	10.9	*9.6*	8.5	9.4	*0.0*	0.0	0.0
	Worst	*23.4*	21.8	22.8	*12.0*	11.3	11.8	*17.4*	14.6	16.3	*0.0*	0.0	0.0
	Nom.	*21.7*	20.6	21.2	*11.2*	10.7	11.0	*9.7*	8.8	9.7	*0.0*	0.0	0.0
P5	Avg.	*25.4*	22.5	22.6	*12.7*	11.2	11.3	*63.3*	62.9	62.6	*1.3*	1.0	1.1
	Worst	*26.3*	23.1	23.2	*13.2*	11.5	11.7	*64.3*	63.7	63.4	*1.7*	1.4	1.5
	Nom.	*25.6*	22.7	22.8	*12.8*	11.3	11.4	*63.0*	62.7	63.2	*1.4*	1.1	1.2

References

[1] A. Elhammali, P. Blanchard, A. Yoder, Z. Liao, X. Zhang, X. R. Zhu, P. K. Allen, M. Jeter, J. Welsh, and Q.-N. Nguyen, Clinical outcomes after intensity-modulated proton therapy with concurrent chemotherapy for inoperable non-small cell lung cancer, Radiotherapy and Oncology **136**, 136–142 (2019).

[2] N. Nakamura, K. Hotta, S. Zenda, H. Baba, S. Kito, T. Akita, A. Motegi, H. Hojo, M. Nakamura, R. V. Parshuram, M. Okumura, and T. Akimoto, Hypofractionated proton beam therapy for centrally located lung cancer, Journal of Medical Imaging and Radiation Oncology **63**, 552–556 (2019).

[3] R. M. Hoshina, T. Matsuura, K. Umegaki, and S. Shimizu, A Literature Review of Proton Beam Therapy for Prostate Cancer in Japan, Journal of Clinical Medicine **8**, 48 (2019).

[4] S. Dowdell, C. Grassberger, G. C. Sharp, and H. Paganetti, Interplay effects in proton scanning for lung: a 4D Monte Carlo study assessing the impact of tumor and beam delivery parameters, Physics in Medicine and Biology **58**, 4137–4156 (2013).

[5] L. Kardar, Y. Li, X. Li, H. Li, W. Cao, J. Y. Chang, L. Liao, R. X. Zhu, N. Sahoo, M. Gillin, Z. Liao, R. Komaki, J. D. Cox, G. Lim, and X. Zhang, Evaluation and mitigation of the interplay effects of intensity modulated proton therapy for lung cancer in a clinical setting, Practical Radiation Oncology **4**, e259–e268 (2014).

[6] C. Grassberger, S. Dowdell, A. Lomax, G. Sharp, J. Shackleford, N. Choi, H. Willers, and H. Paganetti, Motion Interplay as a Function of Patient Parameters and Spot Size in Spot Scanning Proton Therapy for Lung Cancer, International Journal of Radiation Oncology∗Biology∗Physics **86**, 380–386 (2013).

[7] J.-P. Bissonnette, K. N. Franks, T. G. Purdie, D. J. Moseley, J.-J. Sonke, D. A. Jaffray, L. A. Dawson, and A. Bezjak, Quantifying Interfraction and Intrafraction Tumor Motion in Lung Stereotactic Body Radiotherapy Using Respiration-Correlated Cone Beam Computed Tomography, International Journal of Radiation Oncology∗Biology∗Physics **75**, 688–695 (2009).

[8] G. Bosmans, A. van Baardwijk, A. Dekker, M. Öllers, L. Boersma, A. Minken, P. Lambin, and D. D. Ruysscher, Intra-patient variability of tumor volume and tumor motion during conventionally fractionated radiotherapy for locally advanced non-small-cell lung cancer: A prospective clinical study, International Journal of Radiation Oncology∗Biology∗Physics **66**, 748–753 (2006).

[9] H. Paganetti, Range uncertainties in proton therapy and the role of Monte Carlo simulations, Physics in Medicine and Biology **57**, R99–R117 (2012).

[10] S. Brousmiche, K. Souris, J. O. de Xivry, J. A. Lee, B. Macq, and J. Seco, Combined influence of CT random noise and HU-RSP calibration curve nonlinearities on proton range systematic errors, Physics in Medicine & Biology **62**, 8226–8245 (2017).

[11] A. V. Chvetsov and S. L. Paige, The influence of CT image noise on proton range calculation in radiotherapy planning, Physics in Medicine and Biology **55**, N141–N149 (2010).

[12] A. C. Kraan, S. van de Water, D. N. Teguh, A. Al-Mamgani, T. Madden, H. M. Kooy, B. J. Heijmen, and M. S. Hoogeman, Dose Uncertainties in IMPT for Oropharyngeal Cancer in the Presence of Anatomical, Range, and Setup Errors, International Journal of Radiation Oncology∗Biology∗Physics **87**, 888–896 (2013).

[13] K. M. Kraus, E. Heath, and U. Oelfke, Dosimetric consequences of tumour motion due to respiration for a scanned proton beam, Physics in Medicine and Biology **56**, 6563–6581 (2011).

[14] P. C. Park, J. P. Cheung, X. R. Zhu, A. K. Lee, N. Sahoo, S. L. Tucker, W. Liu, H. Li, R. Mohan, L. E. Court, and L. Dong, Statistical Assessment of Proton Treatment Plans Under Setup and Range Uncertainties, International Journal of Radiation Oncology∗Biology∗Physics **86**, 1007–1013 (2013).

[15] A. J. Lomax, Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties, Physics in Medicine and Biology **53**, 1027–1042 (2008).

[16] A. J. Lomax, Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions, Physics in Medicine and Biology **53**, 1043–1056 (2008).

[17] M. van Herk, P. Remeijer, C. Rasch, and J. V. Lebesque, The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy, International Journal of Radiation Oncology∗Biology∗Physics **47**, 1121–1135 (2000).

[18] J. Unkelbach, T. Bortfeld, B. C. Martin, and M. Soukup, Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning, Medical Physics **36**, 149–163 (2008).

[19] A. Fredriksson and R. Bokrantz, The scenario-based generalization of radiation therapy margins, Physics in Medicine and Biology **61**, 2067–2082 (2016).

[20] P. C. Park, X. R. Zhu, A. K. Lee, N. Sahoo, A. D. Melancon, L. Zhang, and L. Dong, A Beam-Specific Planning Target Volume (PTV) Design for Proton Therapy to Account for Setup and Range Uncertainties, International Journal of Radiation Oncology∗Biology∗Physics **82**, e329–e336 (2012).

[21] M. Cubillos-Mesías, M. Baumann, E. G. C. Troost, F. Lohaus, S. Löck, C. Richter, and K. Stützer, Impact of robust treatment planning on single- and multi-field optimized plans for proton beam therapy of unilateral head and neck target volumes, Radiation Oncology **12** (2017).

485 [22] W. Liu, X. Zhang, Y. Li, and R. Mohan, Robust optimization of intensity modulated proton
486 therapy, Medical Physics **39**, 1079–1091 (2012).

487 [23] D. Pflugfelder, J. J. Wilkens, and U. Oelfke, Worst case optimization: a method to account for
488 uncertainties in the optimization of intensity modulated proton therapy, Physics in Medicine
489 and Biology **53**, 1689–1700 (2008).

490 [24] A. Fredriksson, A. Forsgren, and B. Hårdemark, Minimax optimization for handling range and
491 setup uncertainties in proton therapy, Medical Physics **38**, 1672–1684 (2011).

492 [25] J. Unkelbach, T. C. Y. Chan, and T. Bortfeld, Accounting for range uncertainties in the
493 optimization of intensity modulated proton therapy, Physics in Medicine and Biology **52**,
494 2755–2773 (2007).

495 [26] M. Bangert, P. Hennig, and U. Oelfke, Analytical probabilistic modeling for radiation therapy
496 treatment planning, Physics in Medicine and Biology **58**, 5401–5419 (2013).

497 [27] L. V. van Dijk, R. J. H. M. Steenbakkers, B. ten Haken, H. P. van der Laan, A. A. van 't
498 Veld, J. A. Langendijk, and E. W. Korevaar, Robust Intensity Modulated Proton Therapy
499 (IMPT) Increases Estimated Clinical Benefit in Head and Neck Cancer Patients, PLOS ONE
500 **11**, e0152477 (2016).

501 [28] H. Li, X. Zhang, P. Park, W. Liu, J. Chang, Z. Liao, S. Frank, Y. Li, F. Poenisch, R. Mohan,
502 M. Gillin, and R. Zhu, Robust optimization in intensity-modulated proton therapy to account
503 for anatomy changes in lung cancer patients, Radiotherapy and Oncology **114**, 367–372 (2015).

504 [29] W. Liu, Z. Liao, S. E. Schild, Z. Liu, H. Li, Y. Li, P. C. Park, X. Li, J. Stoker, J. Shen,
505 S. Keole, A. Anand, M. Fatyga, L. Dong, N. Sahoo, S. Vora, W. Wong, X. R. Zhu, M. Bues, and
506 R. Mohan, Impact of respiratory motion on worst-case scenario optimized intensity modulated
507 proton therapy for lung cancers, Practical Radiation Oncology **5**, e77–e86 (2015).

508 [30] K. Bernatowicz, X. Geets, A. Barragan, G. Janssens, K. Souris, and E. Sterpin, Feasibility of
509 online IMPT adaptation using fast, automatic and robust dose restoration, Physics in Medicine
510 & Biology **63**, 085018 (2018).

511 [31] D. Cummings, S. Tang, W. Ichter, P. Wang, J. D. Sturgeon, A. K. Lee, and C. Chang, Four-
512 dimensional Plan Optimization for the Treatment of Lung Tumors Using Pencil-beam Scanning
513 Proton Radiotherapy, Cureus (2018).

[32] T. Inoue, J. Widder, L. V. van Dijk, H. Takegawa, M. Koizumi, M. Takashina, K. Usui, C. Kurokawa, S. Sugimoto, A. I. Saito, K. Sasai, A. A. van't Veld, J. A. Langendijk, and E. W. Korevaar, 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, International Journal of Radiation Oncology∗Biology∗Physics **96**, 661–669 (2016).

[33] S. van der Voort, S. van de Water, Z. Perkó, B. Heijmen, D. Lathouwers, and M. Hoogeman, Robustness Recipes for Minimax Robust Optimization in Intensity Modulated Proton Therapy for Oropharyngeal Cancer Patients, International Journal of Radiation Oncology∗Biology∗Physics **95**, 163–170 (2016).

[34] M. Bensimhoun, N-dimensional cumulative function, and other useful facts about gaussians and normal densities, Jerusalem, Israel, Tech. Rep , 1–8 (2009).

[35] OpenReggui - https://openreggui.org/.

[36] J. Amanatides and A. Woo, A Fast Voxel Traversal Algorithm for Ray Tracing, in *In Eurographics 87*, pages 3–10, 1987.

[37] M. Wanet, E. Sterpin, G. Janssens, A. Delor, J. A. Lee, and X. Geets, Validation of the mid-position strategy for lung tumors in helical TomoTherapy, Radiotherapy and Oncology **110**, 529–537 (2014).

[38] MCsquare - http://www.openmcsquare.org/.

[39] K. Souris, J. A. Lee, and E. Sterpin, Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures, Medical Physics **43**, 1700–1712 (2016).

[40] K. Souris, A. B. Montero, G. Janssens, D. D. Perri, E. Sterpin, and J. A. Lee, Technical Note: Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy, Medical Physics (2019).

[41] A. B. Montero, K. Souris, E. Sterpin, and J. Lee, OC-0265: Efficient implementation of random errors in robust optimization for proton therapy with Monte Carlo, Radiotherapy and Oncology **119**, S123–S124 (2016).

[42] A. Fredriksson, A characterization of robust radiation therapy treatment planning methods-from expected value to worst case optimization, Medical Physics **39**, 5169–5181 (2012).