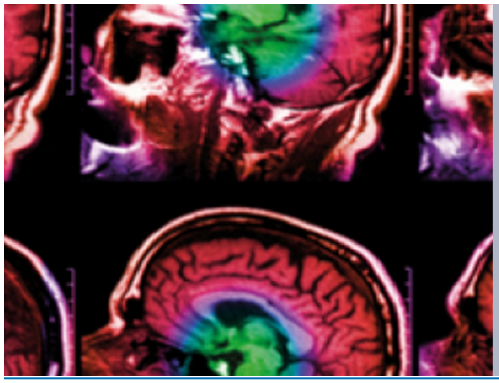
Physics in Medicine & Biology

PAPER

Development of robustness evaluation strategies for enabling statistically consistent reporting



IOP Publishing Phys. Med. Biol. 66 (2021) 045002 https://doi.org/10.1088/1361-6560/abd22f

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
20 May 2020

REVISED
1 December 2020

ACCEPTED FOR PUBLICATION
9 December 2020

PUBLISHED
1 February 2021

PAPER

Development of robustness evaluation strategies for enabling statistically consistent reporting

E Sterpin[1,2], Sara T Rivas[2], F Van den Heuvel[3,4], B George[3], J A Lee[2] and K Souris[2]

[1] KU Leuven, Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium
[2] Université catholique de Louvain, Institut de Recherche Expérimentale et Clinique, Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Brussels, Belgium
[3] CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford, United Kingdom
[4] Dept of Haematology/Oncology, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

E-mail: edmond.sterpin@kuleuven.be

Keywords: proton therapy, robustness evaluation, robust planning, IMPT

Supplementary material for this article is available online

Abstract

Robustness evaluation of proton therapy treatment plans is essential for ensuring safe treatment delivery. However, available evaluation procedures feature a limited exploration of the actual robustness of the plan and generally do not provide confidence levels. This study compared established and more sophisticated robustness evaluation procedures, with quantified confidence levels. We have evaluated several robustness evaluation methods for 5 bilateral head-and-neck patients optimized considering spot scanning delivery and with a conventional CTV-to-PTV margin of 4 mm. Method (1) good practice scenario selection (GPSS) (e.g. +/− 4 mm setup error 3% range uncertainty); (2) statistically sound scenario selection (SSSS) either only on or both on and inside isoprobability hypersurface encompassing 90% of the possible errors; (3) statistically sound dosimetric selection (SSDS). In the last method, the 90% best plans were selected according to either target coverage quantified by D_{95} (SSDS_D_{95}) or to an approximation of the final objective function (OF) used during treatment optimization (SSDS_OF). For all methods, we have considered systematic setup and systematic range errors. A mix of systematic and random setup errors were also simulated for SSDS, but keeping the same conventional margin of 4 mm. All robustness evaluations have been performed using the fast Monte Carlo dose engine MCsquare. Both SSSS strategies yielded on average very similar results. SSSS and GPSS yield comparable values for target coverage (within 0.5 Gy). The most noticeable differences were found for the CTV between GPSS, on the one hand, and SSDS_D_{95} and SSDS_OF, on the other hand (average worst-case D_{98} were 2.8 and 2.0 Gy larger than for GPSS, respectively). Simulating explicitly random errors in SSDS improved almost all DVH metrics. We have observed that the width of DVH-bands and the confidence levels depend on the method chosen to sample the scenarios. Statistically sound estimation of the robustness of the plan in the dosimetric space may provide an improved insight on the actual robustness of the plan for a given confidence level.

1. Introduction

External beam radiotherapy aims at delivering sufficient dose to tumor tissue while preserving surrounding organs. In order to achieve this goal, sophisticated irradiation techniques, such as the use of protons, have been developed in order to conform doses to target volumes. However, radiotherapy treatment delivery can be affected by many sources of uncertainties: patient positioning, inter- and intra-fraction movements, and imperfect conversion of imaging data into physical quantities. In order to secure target coverage and avoid accidental organ-at-risk irradiation, robust planning methods have been developed to ensure that delivered doses keep meeting the objectives and constraints despite uncertainties. In conventional x-ray radiotherapy, this

objective is typically achieved by safety margins with the concept of planning target volume (PTV) and planning risk volume (PRV). In proton therapy, the typical margin strategies suffer from notorious shortcomings, because of the sensitivity of proton therapy dose distributions to the uncertainties of the position of the Bragg peak and failure of the static-dose cloud approximation, which assumes that patient shifts do not change the dose distributions (Albertini *et al* 2011, Fredriksson 2012).

As a result, many methods of robust planning and robustness evaluation have been proposed in the literature for proton therapy (Albertini *et al* 2011, Fredriksson 2012*a*, 2012*b*, Bokrantz and Fredriksson 2013, Casiraghi *et al* 2013, Liu *et al* 2014, Lowe *et al* 2015, Malyapa *et al* 2016). In the case of the most advanced intensity modulated proton therapy (IMPT) techniques, robust planning typically consists of a minimax problem that is solved by optimizing the worst-case scenario among a set of predefined possible scenarios (Fredriksson 2012). Generally, this set of scenarios includes (systematic) positioning errors, image conversion errors, and in some cases the movement of organs represented by additional image sets which are included as additional scenarios (Chang *et al* 2014, De Ruysscher *et al* 2015, Unkelbach *et al* 2018, Ge *et al* 2019). Irrespective of the considered robust planning method, it remains necessary to evaluate the actual robustness of the plan. The evaluation can be done more thoroughly than during robust optimization. Robustness evaluation often includes, for example, random errors, organ deformations, interplay effects, etc. This is because robustness evaluation is computationally less demanding than robust optimization (for instance, no need to store influence matrices). However, the methods typically reported for robustness evaluation are also based on a relatively simple sampling of error scenarios, for example some (systematic) positioning errors combined with image conversion errors. Other authors have also incorporated random errors (Fredriksson 2012, Lowe *et al* 2015).

Most robustness evaluation methods reported in the literature and used in some commercial planning systems may feature several biases because of pragmatic choices imposed by limited computing resources and due to a lack of consensus in the involved concepts. A first bias lies in the direct combination of pre-sampled uncertainties, leading to the selection of very unlikely scenarios, for example setup errors of $+/-5$ mm combined with density errors of $+/-3\%$, i.e. the simultaneous selection of two extremes in the probability distributions. This amounts to combining extremes of marginal probability distributions, while the joint probability distribution should be sampled instead. Korevaar *et al* have already pointed that issue and have performed robustness evaluation using a statistically consistent but limited set of scenarios (Korevaar *et al* 2019). A second bias is the lack of consistently calculated confidence levels, in order to clearly define what is meant by a 'worst-case'. Indeed, the worst-case scenario is the least favorable scenario among a pre-defined selection set (otherwise, the most extreme case can always be envisaged). In the best-known margin calculation recipe, the value of the final margin depends on a choice of the number of patients for which one wishes to ensure target coverage (typically 90%) (van Herk *et al* 2000). This confidence level is not always reported in the literature when it comes to robustness assessments. In addition, a lack of clarity remains on how to calculate this confidence level. Specifically, should it be calculated in the error space, i.e. as the percentage of possible scenarios covered by a given robustness test? Or should it be calculated in the dose space, that is, as the percentage of dose distributions meeting a given clinical endpoint?

In this publication, we compare several robustness evaluation methods, with explicitly calculated confidence levels in either the error space or the dose distribution space.

2. Materials and methods

2.1. Definitions and notations

We define the robustness of a treatment plan as the capability of this plan to continue satisfying clinical objectives and/or constraints despite uncertainties, for a certain confidence level. As in van Herk *et al* (2000), treatment errors can be classified in treatment preparation errors (e.g. systematic errors) and treatment execution errors (e.g. random errors). Like van Herk's formalisation, we suppose knowledge of the probability density functions (*pdf*) of these errors. In general, these are assumed to be normal (Gaussian) with standard deviations Σ and σ for systematic and random errors, respectively. For the remaining of this manuscript, we will limit ourselves to the following errors:

1. Setup errors (se; (x_{se}, y_{se}, z_{se})) characterized by 3D Gaussian *pdfs* for both systematic and random errors, with vector standard deviations $\boldsymbol{\Sigma}_{se}$ and $\boldsymbol{\sigma}_{se}$, respectively.
2. Range uncertainties (RU; due for instance to improper image conversion), characterized by a 1D Gaussian *pdf* with Σ_{RU} as standard deviation.

However, these considerations can be generalized to an arbitrary number of types of uncertainties. When appropriate, the generalization of the developed methods will be addressed.

2.2. Computation of confidence levels

For a given *pdf*, confidence intervals define a range within which a population parameter resides for a given confidence level. In van Herk's margin recipe, a typical confidence level chosen is 90% which leads to the 2.5 factor that multiplies the standard deviation in the well-known formula for 3D-conformal dose distributions: $M_{\mathrm{PTV}} = 2.5\Sigma + 0.7\sigma$. This means that 90% of the possible systematic errors within the patient population will be covered by the margin recipe. However, such a margin recipe fails notoriously in proton therapy because it is based on the static-dose cloud approximation (Stuschke *et al* 2012, Liu *et al* 2013, Liu *et al* 2016). Moreover, the number of fractions is assumed infinite, which allows a simple model to approximate how random errors blur the dose distributions. This simplification leads to the term 0.7σ in the margin recipe, considering a typical 95% of the dose prescription as the minimum dosimetric coverage. A reduced number of fractions requires either a more complex model or the conversion of part of the random error into a systematic component, as acknowledged in van Herk *et al* (2000). Such approach will also not hold in proton therapy because of the failure of the static dose cloud approximation.

Thus, more complex models and formalisms are needed in proton therapy to assess the robustness in lieu of simplistic margin recipes. First, we need to distinguish occurrences of errors and the combined effect of these occurrences (the sum over each fraction) over the entire course of a treatment, referred here as *treatment scenarios* or, shorter, *scenarios*. In the simplified context mentioned here, a scenario will therefore be characterized by a systematic error sampled from a Gaussian distribution with a vector of standard deviations $\boldsymbol{\Sigma}_{\mathbf{se}}$, and a sequence of errors for each treatment fraction that are randomly sampled from a Gaussian distribution with a vector of standard deviations $\boldsymbol{\sigma}_{\boldsymbol{se}}$. Second, it is very unlikely to provide closed-form analytical expressions to characterize how uncertainties affect the dose distributions. Therefore, it is not feasible to derive simple margin recipes with satisfactory mathematical grounds.

In general, a confidence interval for an estimator of interest consists in giving the narrowest range of values for that estimator, such that the *pdf* integrates to 0.9 (90%) over that range. In practice, the *pdf* can be sampled and sorted, after which the suitable bounds can be reported.

2.2.1. Confidence levels in the dosimetric space

A straightforward way to compute a confidence level for a dosimetric estimator is to generate dose distributions for many scenarios and compute the probability that a certain rule on this dosimetric estimator will be realized (for instance, $D_{95} >$ xx Gy with a probability of yy (or confidence)). This will be referred as the computation of a confidence level in the *dosimetric space*. In such approach, we can provide the percentage of times, i.e. the confidence level, that each objective/constraint defined by the radiation oncologist will be satisfied. Another possibility would be to provide a bandwidth for a value of interest and an associated confidence level. For instance, we could provide the range of D_{95} for the CTV, corresponding to the 90% highest D_{95} values. This is a relevant metric to estimate the probability of covering the target as desired. However, this might cause to focus too much on target coverage. In order to provide a fair balance between target coverage and organs-at-risk exposure, another possibility would be to select the best 90% *objective function* (OF) values. The value of the objective function of the accepted plan, with the penalties (/objective function weights) for each organ included in the objective function, provides a good estimate of the clinical compromise accepted by the physicist and the physician at the end of the optimization process. Thus, it provides a quantification of the clinical quality of the plan. Therefore, the classification of the best simulated dose distributions according to the value of their associated objective functions seems ideal from a clinical point of view.

Because the confidence levels are estimated from random sampling of the errors, they will be subject to statistical noise. Therefore, enough scenarios must be simulated for estimating confidence levels with sufficient accuracy. The number of scenarios needed to achieve a given statistical accuracy on the confidence level can be determined using the method developed in Souris *et al* where the statistical uncertainty on the estimated confidence level considered is computed dynamically during the robustness evaluation process (Souris *et al* 2019). The key difficulty resides in the generation of the dose distributions. Fast Monte Carlo dose engines associated with clever statistical stopping criteria (Souris *et al* 2019) or other methods like polynomial chaos expansion (Perkó *et al* 2016) can help for this task.

2.2.2. Confidence levels in the error space

In current practice, robustness evaluation tools are limited to the generation of some occurrences of systematic setup and range errors according to parameters defined by the user. Random errors are typically not simulated. Van Der Voort *et al* have suggested to consider random errors using empirical relations that can convert a combination of systematic and random errors into pure systematic errors (Van Der Voort *et al* 2016). Another method has been suggested by the group of PSI, using a relatively small subset of possible errors, a priori limited

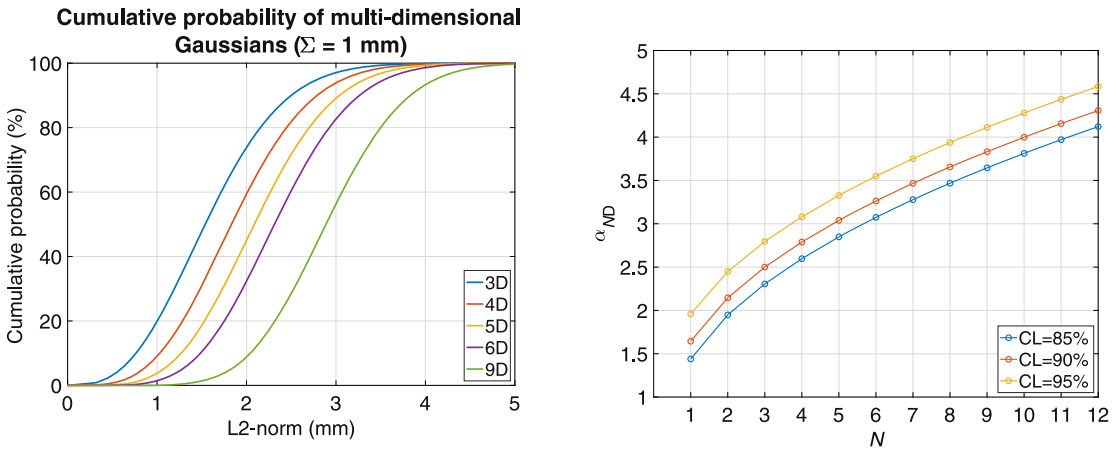


Figure 1. (a) Examples of cumulative probabilities for isotropic multi-dimensional independent normal (Gaussian) distributions with 1 mm standard deviation; (b) values of the reduced radius α_{ND} of the isoprobability hyper-sphere with respect to the number of dimensions for 85, 90, and 95% confidence levels (CL). For 3 dimensions and 90% confidence level (CL), α_{3D} equals the typical 2.5 value found in margin recipe of van Herk *et al* (2000).

by an 85% confidence interval line (Albertini *et al* 2011, Lowe *et al* 2015). For the reminder of the argument, we will assume that random errors are either neglected or converted to systematic errors as in Van Der Voort *et al* (2016).

If dose distributions are unknown, computing confidence levels in proton therapy is not as straightforward as in photon therapy. The main reason is that one cannot easily approximate the effect an error may have on the dose distributions. Consequently, each type of error needs to be considered separately. In the context of independent setup errors and range uncertainties, this leads to the sampling of errors in a 4D space with reduced axis ($x' = \frac{x_{se}}{\Sigma_{setup,x}}$, $y' = \frac{y_{se}}{\Sigma_{setup,y}}$, $z' = \frac{z_{se}}{\Sigma_{setup,z}}$, $\mathrm{RU}' = \frac{\mathrm{RU}}{\Sigma_{\mathrm{RU}}}$), where Σ is the standard deviation. In this space, equiprobable errors will be located on the surface of a hypersphere with equation $x'^2 + y'^2 + z'^2 + \mathrm{RU}'^2 = \alpha_{4D}{}^2$. The parameter α_{4D} denotes the (reduced) radius of the hypersphere. The left side of the last equation represents a chi-square distribution with 4 degrees of freedom. The behavior of the cumulative chi-square distribution is illustrated in figure 1(a) for different numbers of degrees of freedom.

A confidence level in the *error space* can now be approximately computed. To ensure robustness against 90% of all possible scenarios, we need to select all possible configurations within a hypersphere with radius of approximately 2.8 as seen from figure 1. If we hypothesize that the worst-case scenarios are located on the surface of the hyper-sphere, then one can assume that this confidence level of 90% will be achieved by only simulating the points distributed over the hyper-sphere. However, this hypothesis is not necessarily true and will be tested in one of the robustness evaluation strategies introduced in section 2.4.

If range uncertainties are removed, we come back to the 3D case and α_{3D} equals the well-known 2.5 value. Figure 1(b) displays how α_{ND}varies depending on the number of dimensions. It is a direct translation of the value of the L2 norm in figure 1(a) at 90% cumulative probability.

One thing important to note here is that the selection of the scenarios will strongly depend on the dimensionality of the problem. More extreme scenarios will have to be selected for a higher number of dimensions and a fixed confidence level (because of the corresponding increase of the radius α of the hyper-sphere).

2.3. Patient test cases

Five bilateral head-and-neck patients were considered for illustrating the notions described above. Some tumor characteristics are detailed in the appendix (table S1 (available online at stacks.iop.org/PMB/66/045002/mmedia)). The patients were treated by conventional radiotherapy. Hence, the proton treatment plans were optimized for the purpose of this study. The target was the PTV, obtained by expanding the CTV by a 4 mm isotropic margin. The treatment plans included two prescriptions, 70 Gy and 54 Gy on tumor and elective volumes, respectively. The proton treatment plan was composed of 4 scanned beam incidences ((350,60); (350,120);(10,240):(10,300) in degrees for couch and gantry angles, respectively). Treatment plans were optimised to ensure adequate coverage of the PTV, without robustness parameters (treatment plans were not robustly optimized). The minimum requirements were $D_{98} > 90\%$ of prescription dose, $D_{95} > 95\%$ of prescription dose, $D_5 < 105\%$ of prescription dose. However, when possible to respect OAR constraints, we tried to achieve at least 95% of prescribed dose for D_{98}. Constraints to OARs were set according to the clinical

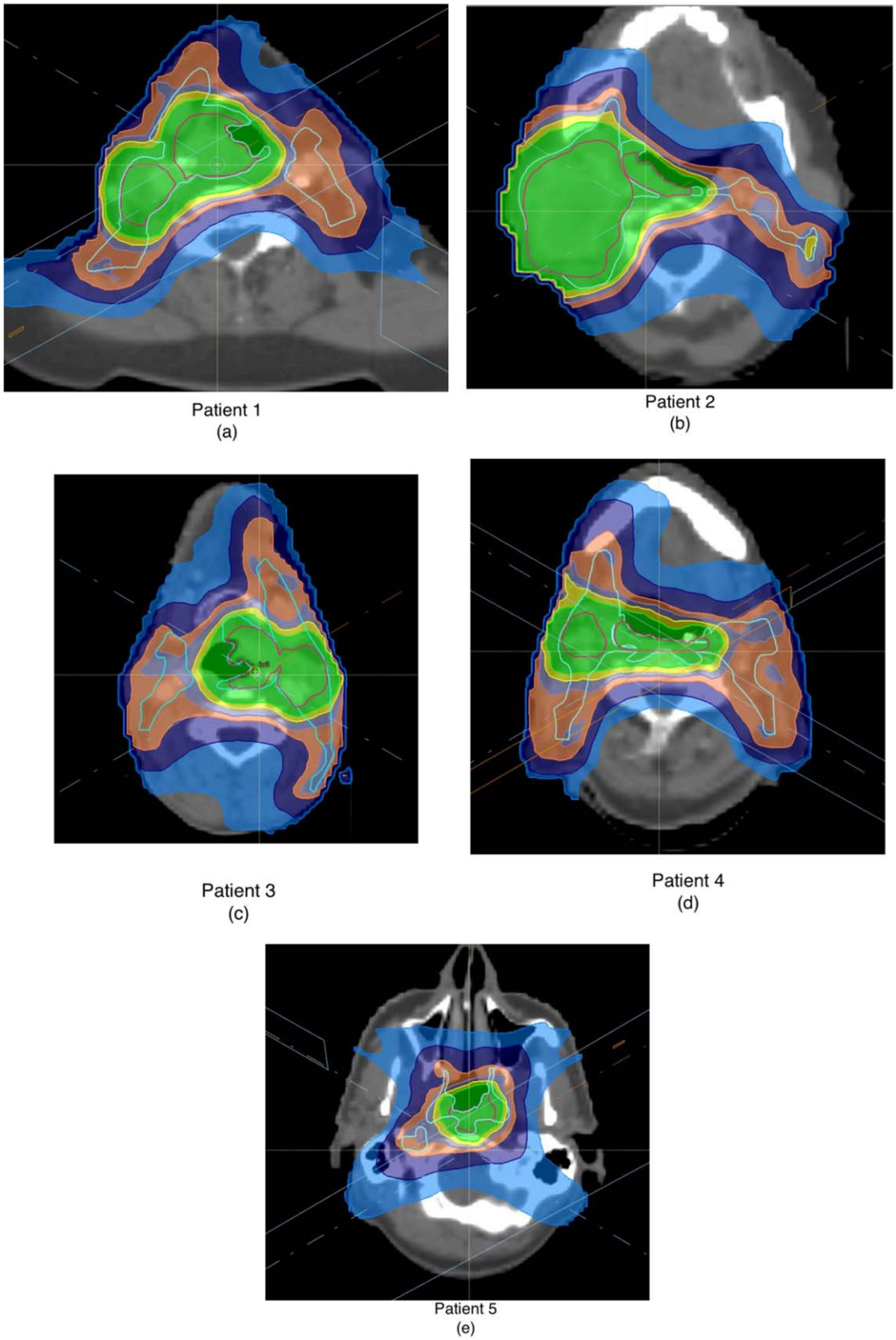


Figure 2. Axial slices of the 5 patients selected for this study with overlaid dose distributions.

rules of Cliniques Universitaires Saint-Luc used for conventional photon therapy. The OARs subject to sparing and their associated dose limits are listed in table S2 in the supplemental material. When possible, the dose to OARs were further diminished provided that it did not compromise PTV coverage. The treatment plans were optimised using RayStation (from RaySearch, research license 5.99). The achieved dose distributions and the used beam angles are illustrated in axial slices for each patient in figure 2.

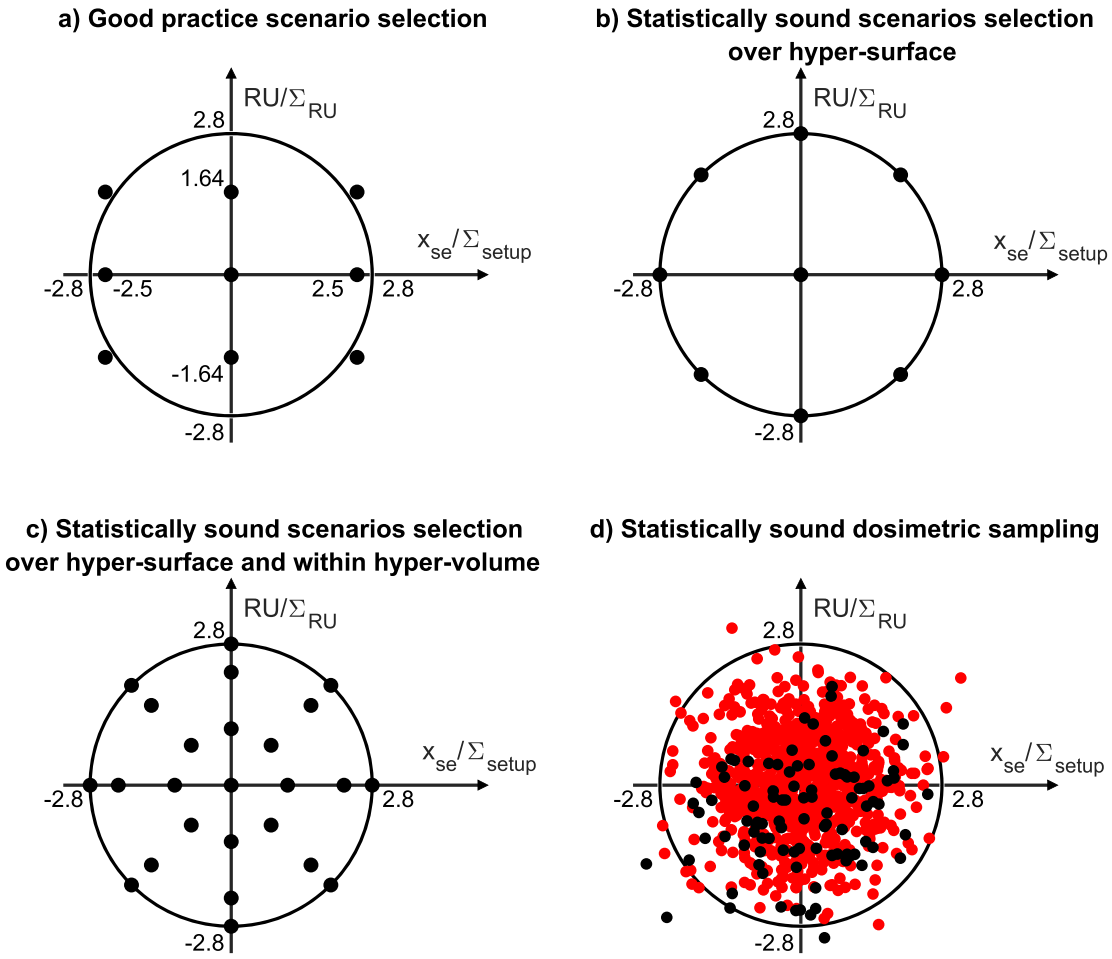


Figure 3. Illustration of the scenario selection methods in a 2D slice (a)–(c), or in a projection (d), of the 4D iso-probability hypersphere. Each dot represents a simulated scenario. The hypersphere contains 90% of all possible scenarios. Figure (a) corresponds to good practice scenario selection (GPSS). Figure (b) corresponds to statistically sound selection of the scenarios (SSSS) at the surface of the isoprobability hyper-sphere including 90% of possible scenarios (SSSS (ON)). Figure (c) corresponds to SSSS at the surface and within the isoprobability hyper-sphere including 90% of possible scenarios (SSSS (IN)). Figure (d) corresponds to statistically sound dosimetric sampling with selection on best 90% D_{95} (SSDS_D_{95}) with a projection of all scenarios onto the plane defined by the *x* and *range* dimensions. Red dots correspond to selected dose distributions; black dots to discarded dose distributions.

The spot positions and weights were exported to a local robustness evaluation tool developed by Souris *et al* (2019). This robustness evaluation tool is based on a validated Monte Carlo dose engine called MCsquare (Souris *et al* 2016). For the purpose of the present study, systematic setup errors and image conversion errors were simulated by shifting the patient and applying a density scaling according to sampled values of setup errors and image conversion errors.

The values chosen for the standard deviations were as follows. For the tests without random errors, $\Sigma_{\text{setup}} = 1.6$ mm, $\sigma_{\text{setup}} = 0$ mm, and $\Sigma_{\text{RU}} = 1.8\%$. The values were chosen in order to represent 4 mm and 3% errors at 90% confidence level in their respective spaces (3D for setup errors ($\alpha_{3D} = 2.5$), 1D for range uncertainties ($\alpha_{4D} = 1.67$)). For the tests with random errors, the values chosen were $\Sigma_{\text{setup}} = 1.3$ mm, $\sigma_{\text{setup}} = 1.0$ mm, and $\Sigma_{\text{RU}} = 1.8\%$. Such combination of systematic and random setup errors leads to a margin of 4 mm using the simplified van Herk formula ($2.5\Sigma + 0.7\sigma$). It is also in line with the empirical relationships shown in figure 3 of Van der Voort *et al* (2016).

2.4. Robustness evaluation strategies investigated

We summarize here the robustness evaluation strategies investigated. A short overview is also given in table 1. In all robustness evaluation strategies, the nominal scenario is kept in the simulated set of dose distributions.

2.4.1. Strategy 1: good practice scenario selection (GPSS) of flat systematic setup and range errors

In many robust optimization/evaluation approaches, scenarios are selected pragmatically according to good practice rules. In general, the CTV to PTV margin is replaced with a systematic setup error of comparable magnitude and the range uncertainty parameter takes typically three values, +RU, 0 and −RU where RU ranges

Table 1. Summary of the robustness evaluation strategies studied and their associated robustness parameters.

Robustness evaluation strategy	Description	Σ_{setup} (mm)	σ_{setup} (mm)	Σ_{RU} (%)
GPSS	Good practice scenario selection in the error space: selection of setup errors onto 90% 3D sphere, and a positive and a negative range value	1.6	0.0	3.0[a]
SSSS (ON)	Statistically sound selection in the error space onto 90% isoprobability line of the 4D hypersphere	1.6	0.0	1.8
SSSS (IN)	Statistically sound selection in the error space onto and inside 90% isoprobability surface of the 4D hypersphere	1.6	0.0	1.8
SSDS_D_{95} (S)	Statistically sound selection in the dosimetric space for the 90% best CTV D_{95}	1.6	0.0	1.8
SSDS_OF (S)	Statistically sound selection in the dosimetric space for the 90% best objective function values	1.6	0.0	1.8
SSDS_D_{95} (R)	Statistically sound selection in the dosimetric space for the 90% best CTV D_{95}	1.3	1.0	1.8
SSDS_OF (R)	Statistically sound selection in the dosimetric space for the 90% best objective function values	1.3	1.0	1.8

[a] For GPSS, only extreme values of the distributions are considered for range errors, not the standard deviations.

from 2.5 to 3.5% in most publications. Random errors are typically ignored or converted into systematic errors, for instance using the approach developed by Van Der Voort *et al* (2016). For this strategy, the setup error and RU parameters equalled 4 mm and 3% (consistent with $\Sigma_{\text{setup}} = 1.6$ mm and $\Sigma_{\text{RU}} = 1.8\%$). In typical clinical practice, only a few scenarios are sampled in the directions *x*, *y*, and *z*, i.e. positive and negative extreme values along each axis (no diagonals). By combining with range errors, it amounts to 20 scenarios in total, excluding the nominal scenario. However, it is not possible with such strategy to estimate a confidence level with acceptable accuracy, as the errors in the spatial directions *x*, *y*, and *z* are sampled too coarsely. Therefore, we have simulated more scenarios by including those on the diagonals between the *x*, *y*, and *z* axes. In such case, the setup errors are selected on the 3D-sphere, at 90% confidence level in 3D (using $\alpha_{3D} = 2.5$). The total amount of scenarios then reaches 80 without the nominal scenario.

In this configuration, a confidence level can be estimated by integrating the joint probability density function inside the 4D hyper-cylinder defined by the 3D setup errors (distributed over a sphere) and the range errors. This was approximated numerically by generating randomly setup and range errors and counting the ones that are inside the hyper-cylinder. This amounts to 81% of possible errors. It is important to mention here that this way of computing the confidence level assume continuity of the errors in the error space and also that the worst errors are located on the edges of the explored space.

For the sake of completeness, we have also simulated the GPSS case with 20 scenarios only. The results are reported in supplementary materials.

2.4.2. Strategy 2: statistically sound scenario selection (SSSS)

Two configurations were tested in this study. In the first configuration, scenarios were sampled uniformly *on* the hyper-surface of the 4D hyper-sphere delimited by the equation $x'^2 + y'^2 + z'^2 + \text{RU}'^2 = {\alpha_{4D}}^2$, where $\alpha_{4D} = 2.8$ to ensure a 90% confidence level in the error space (SSSS (ON) figure 3(b)). In such case, one may assume that this confidence level is secured in the error space provided that robustness for scenarios inside the hyper-surface is also warranted. In the second configuration, scenarios were *also* uniformly sampled *within* the hyper-sphere, in order to better approximate a true 90% confidence level (Perkó *et al* 2016) computed in the error space (figure 3(c)). In SSSS (IN), we also sample hypersurfaces within the 90% hyper-surface with a different radius. The number of scenarios per surface is 80 (3^4 minus the nominal case). In SSSS (IN), we sample 3 hypersurfaces (at (reduced) radii 2.2 and 1.1) hence 240 scenarios. One can note that errors and scenarios lead eventually here to the same meaning, because only systematic errors are sampled.

2.4.3. Strategy 3: statistically sound dosimetric selection (SSDS)

We consider here a Monte Carlo robustness evaluation tool, i.e. errors are randomly sampled according to their *pdfs*. It is worth mentioning that the dose engine associated with this tool can be anything, either Monte Carlo or analytical. A random error sampling approach would be an excellent candidate for performing robustness evaluation because (1) errors can be sampled without any statistical bias from their actual *pdfs*; (2) random errors can be simulated naturally; and (3) it enables an evaluation of the confidence level in the dosimetric space. A weakness of this approach is that the number of treatment scenarios to simulate may be substantial. To ensure its practical viability, dose computation must be performed at a low computational cost. Fast Monte Carlo dose engines may be used for this task, but, in such case, the number of errors and scenarios to simulate must be limited to what is necessary. Therefore, this requires the introduction of a convergence criteria and variance reduction techniques, as described in Souris *et al* (2019). For the purpose of this study, we have tried to minimize the statistical noise as much as reasonably achievable. In Souris *et al* (2019), it was shown that 300 scenarios were sufficient to ensure convergence of the DVH error bands. In the present study, we have therefore simulated 1000 scenarios for ensuring low noise levels on the reported values (for instance, the lowest D_{95} value at 90% confidence level). The number of particles per scenario was about 10^8 to ensure a statistical uncertainty in the target below 2% (one standard deviation). Simulations were performed on a 2x Intel(R) Xeon(R) Gold 6248 CPU.

The SSDS method allows flexibility in the way the scenarios are selected. We implemented two scenario selection methods. In the D_{95} method, the 90% best scenarios according to target coverage for the high dose CTV (quantified here by the D_{95}) were selected for reporting. In the OF method, the 90% best scenarios according to the value of the objective function were selected for reporting. The value of the objective function was computed as a weighted sum of all clinical objectives used in the TPS for the treatment plan optimization. Four objective types, namely minimum dose, maximum dose, mean dose, and DVH objectives, were implemented in the objective function using quadratic terms as described in Oelfke and Bortfeld (2001) (see table S2 in the supplementary material).

Table 2. Metric assessing PTV dose coverage for the nominal plan used in this study. The dose was computed with MCsquare in the nominal case. Target coverage objectives were at least $D_{95} > 95\%$ and $D_{98} > 90\%$ of prescribed dose (70 Gy), thus 66.5 Gy and 63 Gy, respectively. Overdosage were limited by the constraint $D_{95} < 105\%$ (thus 73.5 Gy). When possible, we tried to achieve $D_{98} > 95\%$ of prescribed dose.

PTV metric (Gy)	Patient results for high dose PTV 70 Gy				
	P1	P2	P3	P4	P5
D_{98}	67.6	64.9	67.2	67.9	66.7
D_{95}	68.2	67.0	67.9	68.4	67.5
D_5	71.6	73.4	71.7	72.2	71.6

Table 3. Dose differences between the worst-case and the nominal DVH metrics for the target and organs-at-risk, averaged over the 5 patients (# of scen = number of scenarios; CL = confidence level). The meaning of each robustness evaluation strategy is detailed in table 1. The abbreviation 'prtd' stands for 'parotid'.

Strategy	# of scen	CL (%)	D_{98} CTV 70 Gy (Gy)	D_{95} CTV 70 Gy (Gy)	D_5 CTV 70 Gy (Gy)	D_{98} CTV 54 Gy (Gy)	D_{95} CTV 54 Gy (Gy)	D_{mean} left prtd (Gy)	D_{mean} right prtd (Gy)	D_{mean} oral cavity (Gy)	D_2 spinal cord (Gy)	D_2 brain stem (Gy)
GPSS	80	81	−4.9	−3.9	2.0	−3.9	−2.9	6.4	5.8	3.8	6.7	5.3
SSSS (ON)	80	90	−4.4	−3.6	1.8	−3.5	−2.6	6.1	5.7	3.5	7.0	5.6
SSSS (IN)	240	90	−4.4	−3.6	1.8	−3.5	−2.6	6.1	5.7	3.5	7.0	5.6
SSDS_D_{95} (S)	1000	90	−2.1	−1.3	1.9	−3.9	−3.1	6.7	6.2	4.1	7.0	5.5
SSDS_OF (S)	1000	90	−2.9	−2.2	1.6	−3.5	−2.6	5.3	5.6	3.3	7.0	5.5
SSDS_D_{95} (R)	1000	90	−2.0	−1.1	1.9	−3.4	−2.6	6.1	5.6	3.6	5.8	4.7
SSDS_OF (R)	1000	90	−2.4	−1.8	1.3	−2.8	−2.2	4.8	4.7	2.8	6.2	4.9

For each scenario selection method in the dosimetric space, two tests were performed. In the SSDS (S) strategy, only systematic errors were considered. In the SSDS (R) strategy, both systematic and random errors were considered. The standard-deviations selected for both examples are provided in section 2.3.

3. Results

3.1. Results for the nominal plans

The results obtained for PTV coverage, quantified by the metrics D98, D95 and D5, in the nominal plan using MCsquare are provided in table 2. This computation was necessary to ensure that the dose distributions in the nominal configuration computed by MCsquare met target coverage criteria.

3.2. Comparison of the robustness evaluation methods

Robustness evaluation has been performed for the strategies described in table 1. Table 3 provides the differences between worst-case and the nominal DVH metrics averaged over all patients, for each robustness evaluation strategy. Table 4 displays the same data as table 3, this time with respect to the results yielded by the GPSS method (instead of the nominal plans in table 3). Individual DVH metrics are illustrated for patient 3 in figure 4, and detailed for the same patient in table 5. The same results for the other patients are available in appendix (tables S3–6 and figures S1–4).

For each strategy, the time needed to compute one scenario was about 150 s.

3.2.1. Considering systematic errors only

As shown in table 3 and in the individual results (table 5, figure 4, tables S3–6 and figures S1–4), SSSS (ON) and SSSS (IN) provide very similar DVH metrics. Therefore, they will not be distinguished anymore to present the results. For the high dose CTV, GPSS and SSSS yield similar results, with an average of worst-case D_{98}, D_{95}, and D_5 within 0.5 Gy. Results for individual patients are also similar for the high dose CTV, with differences within 0.7 Gy (tables 5 and S3–6). For the low dose CTV, GPSS and SSSS yield slightly divergent results, with average differences within 0.4 Gy and 0.3 Gy for D_{98} and D_{95}, respectively. The maximum variability occurred for patient 5, with SSSS yielding a D_{98} and a D_{95} 0.9 Gy larger (table S6).

When comparing GPSS to the SSDS methods for target coverage (table 4), differences are more substantial. Considering systematic errors only (S), and D_{98} of the high dose CTV, the worst-case scenario is on average

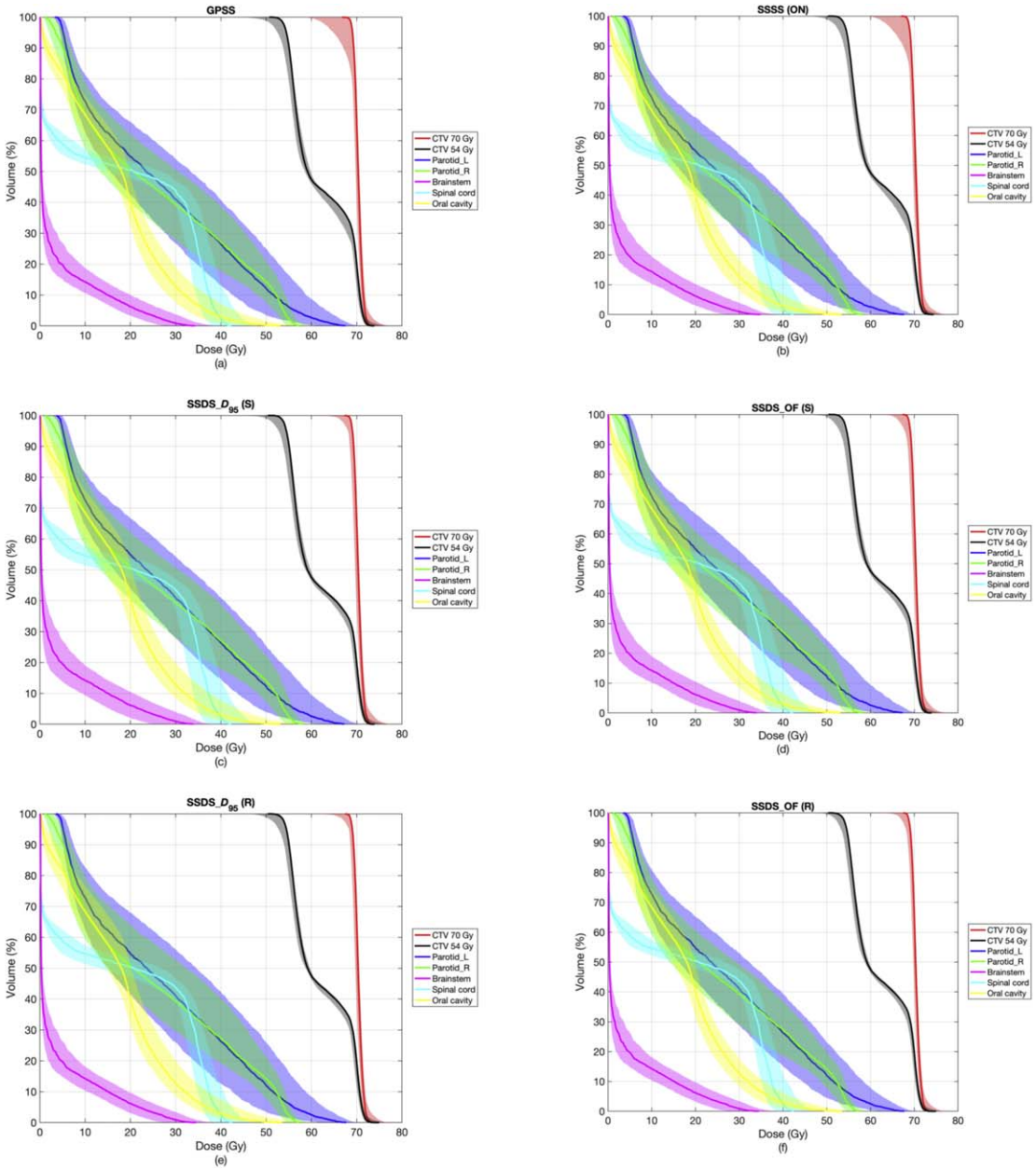


Figure 4. Results of the robustness evaluation for patient 3. The meaning of each robustness evaluation strategy (mentioned in the title of every graph) is detailed in table 1.

Table 4. Average absolute differences of the DVH metrics for the 5 patients with respect to GPSS taken as a reference ('# of scen' = number of scenarios; 'CL' = confidence level). The meaning of each robustness evaluation strategy is detailed in table 1. The abbreviation 'prtd' stands for 'parotid'.

Strategy	# of scen	CL (%)	D_{98} CTV 70 Gy (Gy)	D_{95} CTV 70 Gy (Gy)	D_5 CTV 70 Gy (Gy)	D_{98} CTV 54 Gy (Gy)	D_{95} CTV 54 Gy (Gy)	D_{mean} left prtd (Gy)	D_{mean} right prtd (Gy)	D_{mean} oral cavity (Gy)	D_2 spinal cord (Gy)	D_2 brain stem (Gy)
SSSS (ON)	80	90	0.5	0.3	−0.2	0.4	0.3	−0.3	−0.1	−0.3	0.3	0.3
SSSS (IN)	240	90	0.5	0.3	−0.2	0.4	0.3	−0.3	−0.1	−0.3	0.3	0.3
SSDS_D_{95} (S)	1000	90	2.8	2.6	−0.1	0.0	−0.2	0.3	0.4	0.3	0.3	0.2
SSDS_OF (S)	1000	90	2.0	1.7	−0.4	0.4	0.3	−1.1	−0.2	−0.5	0.3	0.2
SSDS_D_{95} (R)	1000	90	2.9	2.8	−0.1	0.5	0.3	−0.3	−0.2	−0.2	−0.9	−0.6
SSDS_OF (R)	1000	90	2.5	2.1	−0.7	1.1	0.7	−1.6	−1.1	−1.0	−0.5	−0.4

Table 5. Results of the robustness evaluation for patient 3 ('# of scen' = number of scenarios; 'CL' = confidence level). The worst-case are shown for each robustness evaluation strategy. For comparison purposes, the nominal values are also displayed. The meaning of each robustness evaluation strategy is detailed in table 1. The abbreviation 'prtd' stands for 'parotid'.

Robustness evaluation strategy	# of scen	CL (%)	D_{98} CTV 70 Gy (Gy)	D_{95} CTV 70 Gy (Gy)	D_5 CTV 70 Gy (Gy)	D_{98} CTV 54 Gy (Gy)	D_{95} CTV 54 Gy (Gy)	D_{mean} left prtd (Gy)	D_{mean} right prtd (Gy)	D_{mean} oral cavity (Gy)	D_2 spinal cord (Gy)	D_2 brain stem (Gy)
Worst-case												
GPSS	80	81	62.7	64.8	72.4	50.9	52.7	32.4	30.1	19.1	45.0	32.6
SSSS (ON)	80	90	63.1	65.2	72.6	51.1	52.8	32.4	29.8	19.7	45.3	33.1
SSSS (IN)	240	90	63.1	65.2	72.6	51.1	52.8	32.4	29.8	19.7	45.3	33.1
SSDS_D_{95} (S)	1000	90	66.8	68.1	72.6	51.2	52.7	32.9	30.1	20.1	45.4	33.4
SSDS_OF (S)	1000	90	66.3	67.7	72.6	51.2	52.7	32.4	29.9	19.5	45.4	33.4
SSDS_D_{95} (R)	1000	90	67.0	68.3	72.6	51.0	52.7	33.5	30.7	20.4	44.3	32.4
SSDS_OF (R)	1000	90	66.9	68.0	72.4	51.9	53.1	32.1	29.5	19.2	44.9	32.4
Nominal												
	1	NA	68.7	69	71.9	53.8	54.4	26.2	24.9	17	40.2	27.6

2.8 Gy and 2.0 Gy larger for SSDS_D_{95} and SSDS_OF compared to GPSS, respectively. For D_{95}, it is 2.6 and 1.7 Gy, respectively. Comparing GPSS and SSDS_D_{95}, the differences reported are maximum 5.0 Gy and 4.4 Gy higher for D_{98} and D_{95}, respectively (patient 2, table S4). For the low dose CTV, maximum average differences within 0.4 Gy are observed between both SSDS evaluation methods and GPSS. SSSS and SSDS_OF yield on average very similar results for the low dose target (table 4).

These results are confirmed visually in figure 4, where it can be noticed that DVH-bands for the high dose CTV (red) are broader for GPSS and SSSS, than for both SSDS strategies.

For organs-at-risk, the average differences reported are within 1.6 Gy for all metrics between all methods (table 3). It is difficult to distinguish clear trends looking at individual patient results (tables 5 and S3–6). However, one can notice that GPSS often reports the lowest values for OARs. SSDS_OF yields in general similar or lower values than SSSS. Sometimes, SSDS_D_{95} (S) yields substantially larger values than other evaluation methods. For instance, for patient 2, D_{mean} of the left parotid is more than 1.5 Gy larger for SSDS_D_{95} than all other methods (table S4).

3.2.2. Considering systematic and random errors

Simulating explicitly random errors during robustness evaluation yields similar or improved DVH metrics with respect to their counterparts with systematic error only. One can notice in table 3 an average improved coverage of the low dose CTV up to 0.7 Gy for D_{98} (SSDS_OF). For OARs, similar observations can be made, with an improvement of all OAR DVH metrics when random errors are simulated explicitly (i.e. not translated to their approximatively equivalent systematic errors). For instance, the mean to the left and right parotids improved on average in a range from 0.5 Gy to 0.9 Gy.

4. Discussion

The results show that for the patients investigated, SSSS yields the same results whether scenarios are simulated inside the isoprobability sphere or only on the surface. This is in line with previous findings (Casiraghi *et al* 2013, Malyapa *et al* 2016). It is, however, impossible to strictly exclude that a few scenarios inside the hypersphere could lead to unexpected loss of target coverage or unexpected OAR exposure. For instance, range errors induced by setup errors and explicitly simulated range errors could compensate for some particular points on the surface of the hypersphere but not inside, leading to eventually less perturbed dose distributions for some extreme errors. But checking the interior of the hypersphere will inevitably lead to an important increase of the scenarios to simulate (from 80 to 240 in our examples). Therefore, one may consider for practical purposes to explore only the surface of the hypersphere (i.e. the most extreme errors).

A striking result is that GPSS leads to larger error bands for target coverage, often smaller worst-case doses to OARs, and a smaller confidence level of 81%. In practice, this may lead to the decision of replanning because of a lack of target coverage, with the inevitable downside of increasing the dose to OARs, with some that are already slightly underestimated (e.g. 0.3 Gy average difference for D_2 brainstem between GPSS and SSSS). Those results

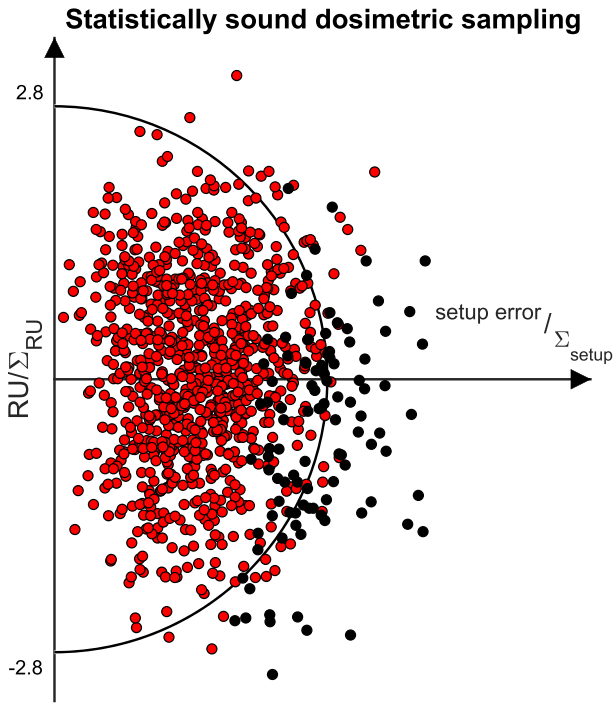


Figure 5. Representation of the scenarios selected by the robustness evaluation SSDS_D_{95}. Each scenario is represented with respect to the simulated range uncertainty and the norm of the sampled setup error (se) $\sqrt{x'^2 + y'^2 + z'^2}$ where $x' = \frac{x_{se}}{\Sigma_{setup}}$, $y' = \frac{y_{se}}{\Sigma_{setup}}$, $z' = \frac{z_{se}}{\Sigma_{setup}}$. Black dots are excluded from the DVH-band while red dots are included in the DVH band. The red dots represent the best 90% values of the D_{95}. The large black circle represents all equiprobable configurations in the (4D) error space (90% confidence level in the error space). Diagram not to scale.

are intuitively expected. Because the GPSS strategy only explores 81% of the possible scenarios (assuming robustness against intermediate errors) AND arbitrarily select extreme scenarios with a very low probability (i.e. outside the 90% hypersphere, thus inconsistent with generally accepted confidence levels (90%)), this leads to an over-conservative approach for the target (because of the extreme cases considered) and a possible under-estimation of the OARs (because of a larger number of unexplored scenarios). An additional source of inconsistency is the arbitrary selection of scenarios with different probabilities (for instance (x_{se}, y_{se}, z_{se}, RU) may equal (4 mm, 0, 0, 0) or (4 mm, 0, 0, 3%) as shown in figure 3(a); the first scenario is more likely to occur). In clinical practice, GPSS is often implemented differently, with a coarser selection of the scenarios in the directions x, y, and z. In such case, the computation of a reliable confidence level becomes very problematic. However, we observe similar results for GPSS either with 20 or 80 scenarios, as it can be seen by comparing tables 3 and S7, which report average results within 0.3 Gy for the targets and 0.8 Gy for the OARs.

The SSSS method will lead overall to the most conservative approach, as shown in tables 3 and 4. Because of the effect of dimensionality (figure 1), SSSS forces the exploration of scenarios that are typically not considered in clinical practice (for photons and protons), for instance an error up to 2.8 Σ_{setup}, which is larger than the more familiar 2.5 Σ_{setup}. The effect of the dimensionality has already been addressed by Korevaar *et al* (2019). If more errors are included, for instance baseline shifts and/or rotations, the errors to explore would be more extreme as shown by figure 4. This is a key weakness of the SSSS method. Because we are blind to the effect of the uncertainties on the dose distributions, the selection can only be performed on or within isoprobability hypersurfaces in order to ensure statistical consistency. As a consequence, the space to explore will increase with the types of errors to explore. In practical cases, the dimensionality of the error space is typically 4D, which leads to a mild increase of the errors to explore (from 2.5 to 2.8 Σ_{setup}). But if a robustness evaluation system aims at improved generalizability of the evaluation, it may need to explore more dimensions (inter-fractional anatomical change, breathing variability, etc), which will inevitably lead to an explosion of the magnitude of the errors and to extremely conservative treatment plans. In the context of the PTV margin recipe, this blindness is overcome by assuming a simple hypothesis related to the dose distributions: the static-dose cloud approximation. This allows a simple sum of the associated random variables—i.e. quadratic sum of variances in margin recipe – so that the problem remains a 3D problem. This hypothesis is rightly forbidden in proton therapy, hence the dimensionality problem that appears here.

The SSDS methods aim precisely at overcoming the downsides of GPSS (inconsistency and arbitrariness) and SSSS (over-conservativeness) discussed above. Because the problem under consideration is eventually a 3D

problem (dose distributions are 3D objects), it is more powerful to explore the scenarios in the dosimetric space. In such case, all the potential redundancies in the error space will be captured. Moreover, extreme errors that may have a low impact on the dose distribution (for instance, a motion parallel to a highly contributing treatment field), can be included in the DVH bands naturally. This can be observed in figure 5, where substantial errors, outside the isoprobability hypersphere, could lead to an acceptable dose distribution. Because what is important in the end is the *confidence level* (i.e. the probability of meeting a criterion or not), statistically unlikely errors can be included safely provided that the final probability (or confidence level) is correctly computed. This leads to a more optimistic estimation of target coverage (2.3 and 1.5 Gy higher on average for D_{98} of the high dose CTV, for SSDS_D_{95} (S) and SSDS_OF (S) compared to SSSS, respectively). And a mild increase (for SSDS_D_{95} or decrease (for SSDS_OF) of DVH metrics of OARs within 0.8 Gy (on average over the 5 patients) compared to SSSS. It is interesting also to mention that such considerations were already addressed for establishing confidence levels for PTV margin recipes. In van Herk *et al* (2000), it is written that 'the margin for treatment preparation (systematic) errors is chosen as a confidence interval that is spherically symmetric. However, an infinite number of 90% confidence intervals may be chosen that are not spherically symmetric. This observation leaves some room for optimization.' In a follow-up paper, Witte *et al* (2017) showed by Monte Carlo simulations how the margin can be optimized to reduce OAR dose while maintaining minimum CTV dose.

However, a new problem that arises is the adequate selection of the scenarios in the dosimetric space. In other words, what is the worst dose distribution? How do we define 'worst'? Tables 3 and 4 show that the reported worst-case will differ significantly depending on the scenario selection method. If we focus on target coverage and select the 90% best D_{95} (SSDS_D_{95}), we obtain the most optimistic result for high dose target coverage, at the expense of generally higher DVH metrics for OARs. Such approach would be ideal in cases with no compromise with respect to OARs. We would then achieve the best estimate of target coverage, for a confidence level of 90%. However, if there are compromises to be made with OARs, then the worst-case dose to OAR will be on the pessimistic side, which may lead to exceed clinical constraints causing the reoptimization of a plan and eventually a deterioration of target coverage.

A solution to the issue of scenario selection based on target coverage only would be to capture the clinical compromise made at the planning level and display the 90% *best* dose distributions, with respect to both target coverage and OAR sparing. We propose here to achieve this by computing for each scenario the *objective function* as accepted by the radiation oncologist and the medical physicist before robustness evaluation. The objective function provides a quantitative assessment of the quality of the plan from a clinical point-of-view, since it integrates clinical objectives and constraints, as well as objective function weights used for optimization that are implicitly approved by the radiation oncologist. Such approach could also naturally be translated to a model-based dose distribution assessment, for instance using tumor control and normal tissue complication probabilities.

The SSDS_OF method yields less optimistic numbers for high dose target coverage than SSDS_D_{95}, but those are still significantly larger than GPSS and SSSS (2.0 and 1.5 Gy larger for D_{98} on average, respectively). However, the results obtained for OARs are on average comparable to both GPSS and SSSS. Interestingly, SSDS_OF also yields results for the low dose target comparable to SSSS. Therefore, SSDS_OF seems to better capture the plans that will lead to the best clinical compromises.

One potential issue of the SSDS_OF method is that objective functions vary by nature from one patient to another depending on the tuning of objective/constraint weights in order to achieve a clinically acceptable compromise between target coverage and OAR sparing. This may lead to undesired variability in robustness reporting. However, such feature could also be seen as an advantage. Two identical robustness evaluation results may lead to different appreciations by a radiation oncologist depending on individual patient characteristics. For instance, more attention can be given to a particular organ-at-risk in a given patient. Such patient-specific characteristics are at least partially entailed implicitly in the objective function. As a consequence, selecting the best dose distributions according to the value of the objective function will tend to be more faithful to the clinical compromises made at the treatment optimization level, and therefore reduce variability in patient reporting from a clinical perspective. Such approach also motivates the radiation oncologist to better formalize the clinical goals he/she aims to achieve before the robust optimization phase starts. This is in line with an improved standardization of the treatment planning workflow, which is essential for its automation.

It is important to note that the computation of confidence levels in the dosimetric space has already been illustrated by Perko *et al* (2016) with the polynomial chaos expansion. Perko *et al* have also identified the potential of working in the dosimetric space to estimate the magnitude of the errors to be included in the robustness evaluation to achieve given statistical criteria, e.g. coverage of the CTV in a given fraction of the patients. The only requirement to work in the dosimetric space is to have a fast dose engine available in order to generate enough dose distributions to compute statistical quantities. This is exactly the purpose of the polynomial chaos expansion method that proposes a novel approach to generate a virtually infinite number of

dose distributions after taking the time to generate a comprehensive dose calculation model (based on about 100 pre-computed scenarios). In our work, we use a fast Monte Carlo dose engine associated with statistically defined stopping criteria to generate the required scenarios. Another difference with the study by Perko *et al* is that the authors evaluate the robustness for each volume of interest separately, while we attempt to evaluate methods to select scenarios globally. The approach of Perko *et al* could be trivially adapted to our methodology. An advantage of a global approach is that it naturally takes into account correlations between the DVH metrics since a set of dose distributions is selected.

The explicit simulation of random errors leads to results that are on average more optimistic than their counterparts with systematic errors only. We remind here that we have always used sets of (Σ, σ) that lead to a consistent CTV-to-PTV margin of 4 mm using the simplified formula of van Herk *et al* ($2.5\Sigma + 0.7\sigma$). This indicates that this formula might be overconservative for the patients investigated in this study. More aggressive plans could therefore be achieved using a statistically sound robustness evaluation method that includes random errors. For instance, SSDS_OF (R) yields a worst-case D_{98} for low dose target that is on average 0.5 Gy higher than SSDS_OF (S). For the right parotid, the worst-case D_{mean} is 0.9 Gy lower for SSDS_OF (R) than SSDS_OF (S). It is interesting to compare SSDS_OF (R) with GPSS by analyzing the last line of table 4. SSDS_OF (R) estimates a better target coverage, overall more optimistic organ-at-risk sparing, and all this for a higher confidence level (90% versus 81%).

It is not the purpose of this paper to suggest a procedure for robustness evaluation. First of all, such procedures will strongly depend on the tumor site considered, the advancement of computing technology, the number of effects we want to consider, and clinical practice. For instance, the group at PSI has suggested a robustness evaluation procedure built up across many publications that is well suited for locations with small systematic errors (Malyapa *et al* 2016). The computation of confidence levels was also included for the effect of fractionation (Lowe *et al* 2015). Other groups have suggested to include variable radiobiological models in their evaluation (Ödén *et al* 2017). However, most robustness evaluation strategies reported in the literature select separately setup errors and range errors according to good practice rules, without considering the computation of confidence levels, neither in the error space nor in the dosimetric space (Liu *et al* 2014, Liu *et al* 2016, van de Water *et al* 2016). As mentioned before, Perko *et al* do compute appropriately confidence levels in the dosimetric space using the polynomial chaos expansion method (Perkó *et al* 2016). Finally, we have reported here worst-case DVH metrics for both target volumes and OARs. One could argue that for parallel-like OAR, like lungs, DVH metrics averaged over the entire set of dose distributions could be more meaningful. In such case, the problem is made trivial for our SSDS methods since we can simply average all DVH metrics over all simulated scenarios. SSSS (IN) should also work. However, adaptations will be required for GPSS and SSSS (ON) since those sample only extreme scenarios, whilst the accurate computation of average DVH metrics would require also intermediate values.

The choice of a robustness evaluation procedure entails also pragmatic considerations such as the time needed to execute the procedure. The SSDS methods are time consuming because enough scenarios need to be simulated in order to minimize the impact of the statistical noise on the reported values. In Souris *et al* (2019), about 300 scenarios seemed adequate to ensure convergence of the results. An intrinsic advantage of Monte Carlo simulations is that the computation time does not scale necessarily with complexity. For instance, random errors can be simulated comprehensively with minimal impact on computation time. Yet, we report here 153 s computation time per scenario, which leads to a total computation time of 13 h for 300 scenarios, which is the maximum limit one may consider in clinical practice (this would correspond to calculations performed overnight). However, such computation time would only be acceptable for a final check, but not for an iterative approach where treatment plans are re-optimized several times according to the results of the robustness evaluation. Therefore, significant improvements are needed to warrant dosimetric selection of scenarios in the clinical practice. This may be achieved by improving the speed of the Monte Carlo dose engine, or the introduction of variance reduction techniques for enabling more efficient sampling of the scenarios, as suggested in Souris *et al* (2019). The polynomial chaos expansion method can also be used to reduce somewhat the number of dose computations needed, and hence speed-up the overall process (Perkó *et al* 2016).

The distinction between the error space and the dosimetric space has been made in the current study for protons only. In general, such distinction is not made in photon therapy because of the usual hypothesis of shift invariance of the dose distributions. If the hypothesis is true, the issue of robustness for target coverage can be formulated as a geometric problem, which leads to safety margin recipes. However, such hypothesis is not necessarily true (for instance, misplaced shoulders in head-and-neck tumors that cause undesired attenuation). Therefore, photon-based treatment plans could also benefit from comprehensive robustness evaluation strategies, which would also help for defining common dose metrics to evaluate proton and photon plans. One can also note that photon-based plans may still benefit from a comprehensive robustness evaluation in the dosimetric space under the hypothesis of shift-invariance of the dose distributions, for instance to reveal

robustness improvements due to non-perfect conformity to the target, or to generate DVH-bands using advanced metrics like the value of the OF.

Finally, it is important to mention that the results presented here were achieved using PTV-based treatment plans, that is, non-robustly optimized. Many papers have shown that robust optimization is more suitable to ensure adequate plan robustness (Unkelbach *et al* 2018). Qualitatively, our conclusions should remain valid if we apply our robustness evaluation methods to robust optimized plans, although this must be confirmed in further studies. Quantitatively, robust optimization is expected to mitigate the differences observed during the present study between the various robustness strategies.

However, complex treatment plans with adjacent target volumes and OARs might lead to challenging clinical trade-offs, even in the context of robust optimization. In such case, having at one's disposal a statistically fair and comprehensive evaluation strategy will help to provide the patients with the best treatment plans, with improved safety. Another limitation of our study resides in the computation of the objective function in the evaluation phase. We have tried to reproduce the best we could the objective function used in the RayStation. However, hidden terms or unforeseen mathematical expressions could be used in the RayStation's objective function and would not be captured by our computation. It would be interesting to compare our results for SSDS_OF to those that would be obtained using the objective function used within the RayStation. Another option would be to design objective functions exclusively for evaluation.

5. Conclusions

Robustness evaluation is a critical step in proton therapy treatment planning. Typically, we aim at evaluating worst-case scenarios within a reasonable set of possible treatment errors. Depending on the outcome of the robustness evaluation, treatment plan optimization may be resumed for enhancing the quality of the plan in terms of target coverage and/or organs-at-risk dose. Therefore, the information delivered by the chosen robustness evaluation strategy must be as accurate and as comprehensive as possible.

We have provided several ways to evaluate statistically the robustness of the plan. An approach based on good practice rules, typically used in current clinical practice, is overall pessimistic for target coverage and optimistic for organs-at-risk sparing, with a relatively low confidence level (81%). Exploring the possible scenarios in the error space in a statistically consistent fashion enables a larger and more familiar confidence level (90%), but at the cost of conservative evaluations of worst-case DVH metrics.

Another approach would be to select scenarios in the dosimetric space, i.e. to select the best dose distributions according to *a priori* defined clinical criteria. Focusing on target coverage provides considerably more optimistic target coverage metrics (and mildly pessimistic OAR sparing). This would probably be a good approach when OAR sparing is easily achievable, and one wants to deliver the most conformal dose possible to achieve target coverage for a given confidence level. A more balanced approach would be to classify the best dose distributions according to the value of the objective function accepted by the radiation oncologist. In such case, a good balance is obtained between the reported worst-case target coverage and OAR sparing. Such approach could be easily implemented in existing commercial solutions.

Acknowledgments

This work is partially inspired from discussions within the European Particle Therapy Network working group 5. Kevin Souris is funded by the Walloon region (MECATECH/BIOWIN, grant number 8090). Sara T Rivas is supported by the Walloon region ('Convention hors pôles ProTherWal', grant number 7289). J. A. Lee is a Senior Research Associate with the Belgian fund of scientific research (F.R.S.-FNRS). Ben George is supported by a Cancer Research UK Centres Network Accelerator Award Grant (A21993) to the ART-NET consortium. Edmond Sterpin's research is partially supported by 'Fonds Baillet-Latour'.

References

Albertini F, Hug E B and Lomax A J 2011 Is it necessary to plan with safety margins for actively scanned proton therapy? *Phys. Med. Biol.* **56** 4399–413

Bokrantz R and Fredriksson A 2013 *Controlling Robustness and Conservativeness in Multicriteria Intensity-Modulated Proton Therapy Optimization Under Uncertainty* Trita-MAT. OS

Casiraghi M, Albertini F and Lomax A J 2013 Advantages and limitations of the 'worst case scenario' approach in IMPT treatment planning *Phys. Med. Biol.* **58** 1323–39

Chang J Y *et al* 2014 Clinical implementation of intensity modulated proton therapy for thoracic malignancies *Int. J. Radiat. Oncol. Biol. Phys.* **90** 809–18

De Ruysscher D, Sterpin E, Haustermans K and Depuydt T 2015 Tumour movement in proton therapy: solutions and remaining questions: a review *Cancers (Basel)* **7** 1143–53

Fredriksson A 2012a A characterization of robust radiation therapy treatment planning methods—from expected value to worst case optimization *Med. Phys.* **39** 5169–81

Fredriksson A 2012b Automated improvement of radiation therapy treatment plans by optimization under reference dose constraints *Phys. Med. Biol.* **57** 7799–811

Ge S *et al* 2019 Potential for improvements in robustness and optimality of intensity-modulated proton therapy for lung cancer with 4-dimensional robust optimization *Cancers* **11** 35

Korevaar E W *et al* 2019 Practical robustness evaluation in radiotherapy – a photon and proton-proof alternative to PTV-based plan evaluation *Radiother. Oncol.* **141** 267–74

Liu W *et al* 2013 Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers *Med. Phys.* **40** 051711

Liu W *et al* 2014 Robust optimization of intensity modulated proton therapy *Med. Phys.* **39** 1079–91

Liu W *et al* 2016 Exploratory study of 4D versus 3D robust optimization in intensity modulated proton therapy for lung cancer *Int. J. Radiat. Oncol. Biol. Phys.* **95** 523–33

Lowe M, Albertini F, Aitkenhead A, Lomax A J and Mackay R I 2015 Incorporating the effect of fractionation in the evaluation of proton plan robustness to setup errors *Phys. Med. Biol.* **61** 413–29

Malyapa R, Lowe M, Bolsi A, Lomax A J, Weber D C and Albertini F 2016 Evaluation of robustness to setup and range uncertainties for head and neck patients treated with pencil beam scanning proton therapy *Int. J. Radiat. Oncol. Biol. Phys.* **95** 154–62

Ödén J, Eriksson K and Toma-Dasu I 2017 Incorporation of relative biological effectiveness uncertainties into proton plan robustness evaluation *Acta Oncol. (Madr)* **56** 769–78

Oelfke U and Bortfeld T 2001 Inverse planning for photon and proton beams *Med. Dosim.* **26** 113–24

Perkó Z, Van Der Voort S R, Van De Water S, Hartman C M H, Hoogeman M and Lathouwers D 2016 Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion *Phys. Med. Biol.* **61** 4646–64

Souris K, Barragan Montero A, Janssens G, Di Perri D, Sterpin E and Lee J A 2019 Technical note: Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy *Med. Phys.* **46** 4676–84

Souris K, Lee J A and Sterpin E 2016 Fast multi-purpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures *Med. Phys.* **1700** 1–23

Stuschke M, Kaiser A, Pöttgen C, Lübcke W and Farr J 2012 Potentials of robust intensity modulated scanning proton plans for locally advanced lung cancer in comparison to intensity modulated photon plans *Radiother. Oncol.* **104** 45–51

Unkelbach J *et al* 2018 Robust radiotherapy planning *Phys. Med. Biol.* **63** 22TR02

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

van de Water S, van Dam I, Schaart D R, Al-Mamgani A, Heijmen B J M and Hoogeman M S 2016 The price of robustness; impact of worst-case optimization on organ-at-risk dose and complication probability in intensity-modulated proton therapy for oropharyngeal cancer patients *Radiother. Oncol.* **120** 56–62

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

Witte M G, Sonke J J, Siebers J, Deasy J O and Van Herk M 2017 Beyond the margin recipe: the probability of correct target dosage and tumor control in the presence of a dose limiting structure *Phys. Med. Biol.* **62** 7874–88