How should we model and evaluate breathing interplay effects in IMPT?

Pastor-Serrano, Oscar; Habraken, Steven; Lathouwers, Danny; Hoogeman, Mischa; Schaart, Dennis; Perkó, Zoltán

DOI
10.1088/1361-6560/ac383f

Publication date
2021

Document Version
Final published version

Published in
Physics in Medicine and Biology

Physics in Medicine & Biology

PAPER • OPEN ACCESS

How should we model and evaluate breathing interplay effects in IMPT?

To cite this article: Oscar Pastor-Serrano *et al* 2021 *Phys. Med. Biol.* **66** 235003

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

OPEN ACCESS

RECEIVED
11 October 2021

ACCEPTED FOR PUBLICATION
10 November 2021

PUBLISHED
23 November 2021

PAPER

How should we model and evaluate breathing interplay effects in IMPT?

Oscar Pastor-Serrano[1,*], **Steven Habraken**[2,3], **Danny Lathouwers**[1], **Mischa Hoogeman**[2,3], **Dennis Schaart**[1,3] and **Zoltán Perkó**[1]

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

E-mail: o.pastorserrano@tudeft.nl

Keywords: Intensity Modulated Proton Therapy (IMPT), breathing motion, statistical evaluation, 4DCT robust optimization, ITV robust optimization, breathing interplay effects

Abstract

Breathing interplay effects in Intensity Modulated Proton Therapy (IMPT) arise from the interaction between target motion and the scanning beam. Assessing the detrimental effect of interplay and the clinical robustness of several mitigation techniques requires statistical evaluation procedures that take into account the variability of breathing during dose delivery. In this study, we present such a statistical method to model intra-fraction respiratory motion based on breathing signals and assess clinical relevant aspects related to the practical evaluation of interplay in IMPT such as how to model irregular breathing, how small breathing changes affect the final dose distribution, and what is the statistical power (number of different scenarios) required for trustworthy quantification of interplay effects. First, two data-driven methodologies to generate artificial patient-specific breathing signals are compared: a simple sinusoidal model, and a precise probabilistic deep learning model generating very realistic samples of patient breathing. Second, we investigate the highly fluctuating relationship between interplay doses and breathing parameters, showing that small changes in breathing period result in large local variations in the dose. Our results indicate that using a limited number of samples to calculate interplay statistics introduces a bigger error than using simple sinusoidal models based on patient parameters or disregarding breathing hysteresis during the evaluation. We illustrate the power of the presented statistical method by analyzing interplay robustness of 4DCT and Internal Target Volume (ITV) treatment plans for a 8 lung cancer patients, showing that, unlike 4DCT plans, even 33 fraction ITV plans systematically fail to fulfill robustness requirements.

1. Introduction

In Intensity Modulated Proton Therapy (IMPT), breathing interplay effects arise from the interaction between the scanning beam and moving organs during treatment delivery. This is detrimental, as during the few minutes in which each fraction is delivered, the continuous movement of the target due to breathing degrades the final dose distribution (Lambert *et al* 2005, Bert *et al* 2008, Bert and Durante 2011). Given the adoption of IMPT in treating moving tumors, there is a growing need for computational methods that allow sound statistical evaluation of interplay effects, where the error introduced by modeling approximations (e.g. using few breathing realizations of sinusoidal breathing) is known and justified.

Several techniques aimed at minimizing the detrimental effect of breathing during delivery include beam gating, rescanning, beam tracking, breath-hold and compression. During beam gating the patient breathes freely and the dose delivery is constrained to a specific part of the breathing cycle (e.g. end of exhale) (Ohara *et al* 1989, Bert *et al* 2009). Beam tracking consists of adjusting the treatment delivery system to real-time predicted target movement (Bert *et al* 2007, Zhang *et al* 2014). In rescanning or repainting the target is irradiated several times

during the same fraction, which helps smooth the final dose distribution (Phillips *et al* 1992, Seco *et al* 2009). Finally, breath-hold and compression methods aim at immobilizing the target during delivery (Boda-Heggemann *et al* 2016, Pguret *et al* 2016).

From a treatment planning perspective, different approaches are used to account for target motion by including information about different breathing phases (e.g. exhale, inhale, mid-ventilation) into the optimization. Internal Target Volume (ITV) planning aims at irradiating an ITV volume in the reference phase, which is defined as the union of all Clinical Target Volume (CTV) contours of the different breathing phases (Shih *et al* 2004). With the help of surrogate models that generate artificial motion, ITVs can be extended to form probabilistic ITVs that capture breathing variability (Krieger *et al* 2020). 4DCT planning is based on optimizing the dose distribution using minimax robust optimization (Pflugfelder *et al* 2008), including multiple Computerized Tomography (CTs) from different breathing phases so that the dose prescriptions are met in all the included breathing phases (Engelsman *et al* 2006, Heath *et al* 2009, Bernatowicz *et al* 2017). Some 4DCT approaches also account for beam tracking (Eley *et al* 2014) and temporal structure (Engwall *et al* 2018b) during optimization.

Interplay effects in IMRT versus IMPT

Previous work shows that fractionation effectively limits interplay dose degradation in Intensity Modulated Radiation Therapy (IMRT) delivery techniques with moving parts such as multi-leaf collimators (MLCs) (Bortfeld *et al* 2002, Jiang *et al* 2003), but may be insufficient to tackle negative biological effects in treatments with many segments of few monitor units (MUs) (Seco *et al* 2007). Several studies further investigate the effects of regular breathing motion and collimator speed on the outcome of MLC treatments (Court *et al* 2008, 2010), showing that non-negligible interplay effects increase with target magnitude, plan complexity and breathing period.

While the problem of interplay is common for all dynamically delivered treatments, its nature differs between IMPT and IMRT: proton pencil beams are narrower than photon fields, deliver the dose more locally, and their irradiation times are usually an order of magnitude smaller. Several studies quantify the negative effect of interplay in IMPT and evaluate the effectiveness of different mitigation techniques such as repainting in lung and liver patients (Seco *et al* 2009, Li *et al* 2014, Zhang *et al* 2016, Engwall *et al* 2018a), breath-hold (Yu *et al* 2017, Emert *et al* 2021) or a comparison between different mitigation techniques used in liver treatments (Zhang *et al* 2018), showing that neither rescanning nor gating alone can mitigate interplay effects. Regarding the effect of motion parameters, large breathing amplitudes are known to produce significant local under- and overdosing (Kraus *et al* 2011, Kardar *et al* 2014, Jakobi *et al* 2018).

Challenges in the evaluation of interplay

Evaluating interplay is usually time consuming and requires many dose distributions corresponding to different realizations of breathing during treatment delivery. While alternative, more realistic and computationally demanding approaches use simulated 4DCT scans with dynamic dose delivery (Boye *et al* 2013) or motion surrogates (den Boer *et al* 2021), most of the interplay evaluation studies are based on a single 4DCT scan and many breathing signals to simulate different breathing scenarios. Obtaining enough of such signals involves either taking fragments from the recorded respiratory signal—which is often short and does not offer much variability—or using a sinusoidal approximation, oversimplifying breathing and failing to capture typical irregularities such as baseline shifts and amplitude changes. Furthermore it is not known how realistic and irregular these signals need to be, how small breathing variations affect the final dose, and how many different breathing samples are needed to accurately capture the statistical variation of interplay. Except for one published paper hinting the possible systematic error in IMRT interplay evaluation caused by the use of a limited number of motion samples in both planning and evaluation (Evans *et al* 2005), no previous study has investigated the statistical significance of evaluating interplay effects using few samples and simplified breathing models disregarding any breathing cycle hysteresis.

Contributions

Building on previous IMRT studies (Kissick *et al* 2005, Seco *et al* 2007), we investigate the interplay dependence on breathing uncertainties for proton treatments with many pencil beams—where the order of magnitude of beam delivery times is a factor 100 lower that the period of breathing motion—and specifically the relationship between dose and breathing parameters such as period and amplitude changes. We also extend on previous work (Court *et al* 2008, 2010) and evaluate interplay using both constant and variable breathing periods. Our analysis is based on a 4DCT scan representing the different anatomies of the patient in a breathing cycle, and breathing signals that capture how these alternate during the course of a treatment fraction. The contributions of this paper are the following:

- We present a method to statistically assess interplay effects in lung IMPT based on breathing signals and apply it to evaluate robustness, comparing the 4DCT and ITV planning approaches and the impact of fractionation for 8 stage III lung cancer patients.
- We evaluate the error in the interplay evaluation caused by (i) using simplistic sinusoidal breathing approximations, (ii) using a limited set of scenarios, or (iii) disregarding breathing hysteresis.
  - (i) Two methods to generate patient-specific breathing signals which differ in accuracy and computational complexity—referred to as *breathing models*—are compared. Specifically, given the popularity of sinusoidal models, we investigate the dosimetric impact of evaluating motion using simple sinusoidal breathing patterns, which is the most commonly used approach when lacking a sufficiently long recorded breathing signal with enough variation.
  - (ii) We investigate the inaccuracies that arise in the statistical evaluation of interplay effects when the analyses lack statistical power (i.e. only a consider a limited number of breathing scenarios), and whether evaluating interplay with a small number of such breathing scenarios—as observed in most of the previous studies—leads to significant errors.
  - (iii) We assess the dosimetric impact of disregarding hysteresis in the breathing cycle, which translates into considering symmetrical inhale and exhale during evaluation.
- We investigate the dependence of IMPT interplay dose distributions on the breathing parameters such as amplitude, period or starting phase. More specifically, we investigate how the dose and Dose Volume Histogram (DVH) values vary with small changes in the breathing signal, and study which parameters (e.g. breathing amplitude, period) have the biggest effect on dose.

The rest of the paper is structured as follows: in section 2, we describe the patient data, the proposed methodology to simulate breathing interplay effects, and explain the design choices in treatment planning and delivery simulation, together with a description of our statistical evaluation. In section 3 we present the results of the numerical experiments, followed by an analysis and discussion in section 4. Finally, we provide final remarks and conclusions in section 5.

2. Methods and materials

Patient data and treatment plans

Different breathing signals are obtained with the stereotactic body radiation surgery (SBRT) system Cyberknife® (Accuray Inc., Sunnyvale CA, US), which tracks targets that move with respiration using a correlation model that relates the internal target position with external markers taped to the chest of the patient (Coste-Manire *et al* 2005, Hoogeman *et al* 2009). The long respiratory traces represent tumor movement during treatment for 8 different lung cancer patients. Each signal is matched to a 4DCT scan from a stage III lung cancer patient (having been treated with IMRT and recorded with a Siemens Sensation Open® CT scanner using phase binning) and subsequently rescaled to the maximum 4DCT amplitude. The 4DCT scans are discretized into 8 phases in the breathing cycle: 0%, 25%, 50%, 75%, 100% inhale, and 75%, 50% and 25% exhale. The structures of interest are clinically delineated in all scans, with the exception of the ITV, which is obtained by combining in the mid-ventilation 50% exhale reference phase the CTV volumes from all the breathing phases. Table 1 describes the motion and tumor sizes of the patients in the dataset. We use ITV plans targeting the ITV in the reference phase, and 4DCT robust plans targeting CTV contours from three phases: the reference 50% exhale phase, and the two extreme 0% and 100% phases. Both ITV and 4DCT robust IMPT plans use a 5 mm setup robustness setting, a 5% range robustness setting, and a 2 mm extra margin around the target(s), based on current clinical practice at Holland PTC (Delft, Netherlands). The treatment is divided into 33 fractions of 2 Gray (Gy), with plans made using Erasmus-iCycle, an in-house Treatment Planning System (TPS) which uses automated multi-criteri a prioritized optimization and a pencil beam dose algorithm to calculate the dose delivered per spot (Breedveld *et al* 2012, van de Water *et al* 2013), including range shifters and filtering of low-weight beams. No breathing uncertainty mitigation technique is applied during planning or delivery, except for one experiment where we apply volumetric repainting per field.

Interplay dose calculation

The proposed model calculates an interplay dose distribution based on the treatment plan, the machine parameters, a 4DCT scan and a breathing signal that can be either a fragment of the real recorded signal or an artificial signal from one of the breathing models discussed below. The number of spots—regions irradiated by a

Table 1. Dataset description and treatment delivery times. The reported values include the breathing amplitude along the lateral, anterior–posterior (A–P) and cranial–caudal (C–C) axes, and the combined volume of the CTV including lymph nodes. We include the treatment delivery time per field for both the 4DCT and ITV plans.

Patient	Breathing amplitude (mm)			Target size (cm^3)	Delivery time (s)					
					Field 1		Field 2		Field 3	
	Lateral	A–P	C–C	CTV and nodes	4DCT	ITV	4DCT	ITV	4DCT	ITV
1	3.2	1.8	4.1	39.1	46.8	46.1	58.2	57.9	28.3	29.1
2	2.1	2.4	5.9	130.9	38.4	33.0	67.9	60.2	39.4	34.8
3	1.5	3.6	9.4	211.5	33.6	29.7	66.2	61.3	35.6	28.5
4	1	0.7	8.7	489.7	68.0	62.2	101.0	96.0	76.5	56.8
5	0.8	0.4	2.4	400.6	34.0	31.9	98.3	76.5	50.4	42.3
6	0.8	2.4	1.7	286.7	53.3	52.6	87.9	65.5	49.7	42.8
7	1.2	1.9	5.4	404.5	61.5	61.7	76.0	76.5	40.2	39.6
8	1.2	0.3	2.3	162.1	40.6	38.5	90.5	69.9	65.5	44.2

single mono-energetic pencil beam—and the order in which they are delivered can be obtained from the treatment plan. Spots are ordered in descending order according to pencil beam energies, on a per field basis. The machine parameters determine a *spot-timeline*, which is a list ordering the spots in time using information such as the elapsed time between two consecutive spots or the time needed to change layers and beams. The irradiation time for each spot is directly obtained from the optimized plans with beam data corresponding to standard Varian ProBeam® settings, resulting in a fixed current and variable local dose rates between 10 and 54 Gy s^{-1}. Beam data measurements are based on integral dose depth curves, lateral spot profiles and absolute dosimetry (MU calibration) under reference conditions in a water phantom. For the machine parameters, 10 ms off-beam time are added after delivery of each spot, as well as an average of 0.7 s to change energy layer. Range shifter fixed insertion times equal 16 s, while the variable time needed to change the gantry angle depends on a linearly increasing, bounded angular acceleration. Figure 1 illustrates the process of simulating an interplay dose distribution. After a breathing signal is generated and a treatment starting phase is sampled, the signal is binned between full inhale and exhale, according to the maximum 4DCT amplitude (5 bins delimited by 4 red horizontal boundary lines in figure 1). The breathing signal indicates the phase in which each spot is delivered, with all the points of the signal that fall between consequent binning boundaries being considered to be the part of the same being phase. For fractions where the patient presents shallow breathing with low amplitude, the dose will be deposited in only a subset of phases. For baseline shifts, the dose delivery will gradually shift from inhale (75%In, 100%, 75%Ex) to exhale phases (25%Ex, 0%, 25%In) as the treatment proceeds. The number of phases used during interplay evaluation may differ from the number of phases used to optimize the 4DCT plan. In this study, 4DCT plans are made using 3 phases, whereas all the 8 available phases are used for the evaluation. After binning, each point of the signal corresponds to a phase of the 4DCT, resulting in the *CT-timeline* containing the different phases ordered in time. Pairing the CT- and spot-timelines results in each spot being assigned to a certain phase. Dose distributions per phase are obtained by adding the doses from individual spots in the same phase, which are later transformed (via a non-rigid thin-plate spline registration deformation field) to the reference phase before being added to form the final dose distribution.

Breathing models

Breathing signals are used to represent respiratory motion during a treatment fraction, and each of them ultimately results in a different dose distribution. The statistical evaluation of interplay effects requires statistics of the dosimetric quantities of interest using many different dose distributions, requiring a large set of respiratory traces. Except for this study where we use breathing signals deliberately recorded during a long time, the available signals from regular patients are usually short and do not contain enough variability, thus requiring commonly used artificial sinusoidal approximations that potentially introduce errors. We compare two different types of data-driven breathing models that capture uncertainty and variability in respiratory motion. The first model is based on simple sinusoidal waves (denoted as 'sin' in the remainder of the paper), while the second model is based on the Adversarial Autoencoder (AAE) algorithm described in Pastor-Serrano *et al* (2021).

1. Sinusoidal model. In the sinusoidal model the respiratory time series is generated by using a sinusoidal function $\sin^{2n}$ as $x(t) = A_0 + A \cdot \sin^{2n}(\pi t/T + \psi)$, where $x(t)$ is the time dependent position of the tumor, A_0 is the position at the beginning of inhale (in centimeters), T is the breathing period (time between two consecutive inhales, in seconds), and A is the amplitude (distance from inhale to exhale, in centimeters).

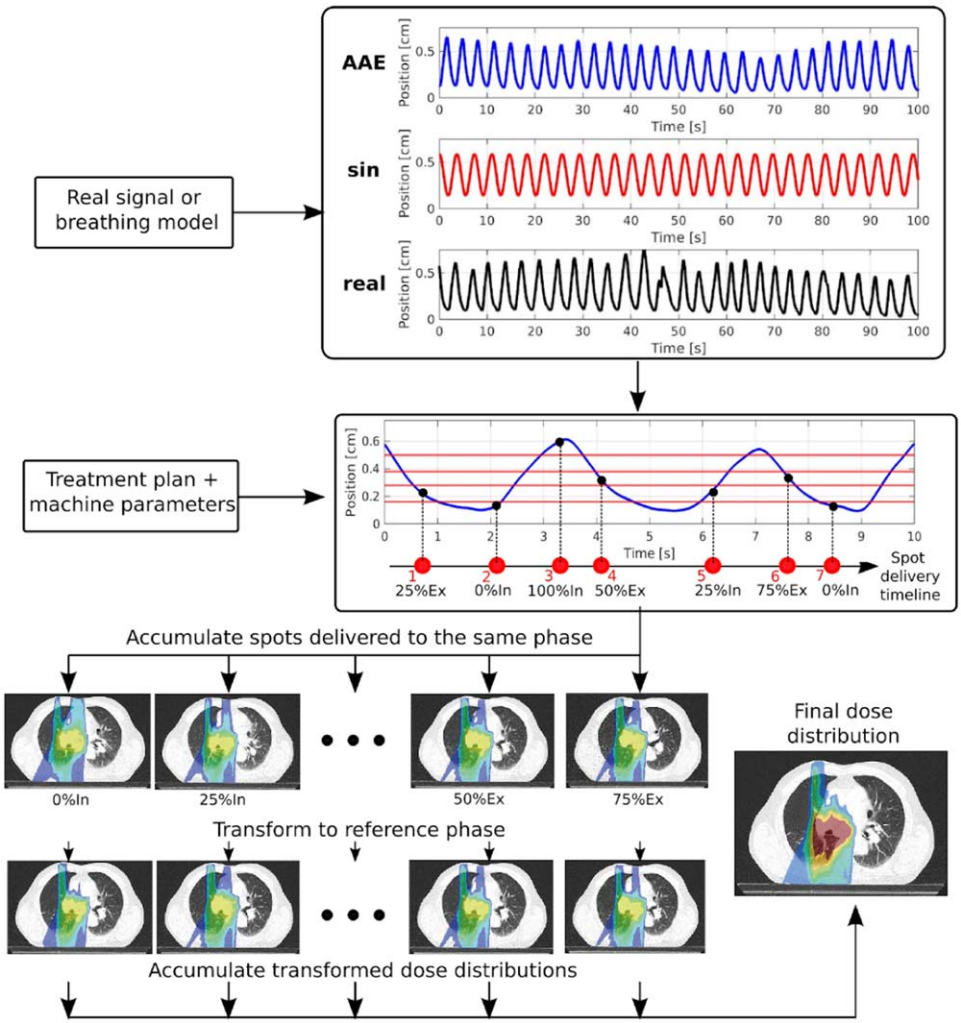


Figure 1. Interplay calculation workflow. The input breathing signal, treatment plan and machine parameters are used to distribute the spots over the breathing phases included in the 4DCT scan and determine in which phase each spot is delivered. Breathing phase dose distributions are first obtained by accumulating the dose delivered by all the spots in the same phase, and are subsequently transformed to the reference phase. The final interplay dose distribution is the result of adding the transformed doses.

The parameter ψ represents offset in phase, and effectively symbolizes the moment when the treatment starts within the first cycle. In our study, we consider the simplest sinusoidals $\sin^{2n}$ with $n = 1$ (Lujan *et al* 2003), constant amplitude and period. Each signal has a different period and amplitude sampled from Gaussian distributions fitted to both the periods $\mathcal{N}(\mu_T, \sigma_T)$ and amplitudes $\mathcal{N}(\mu_A, \sigma_A)$ present in the recorded breathing signal. The parameter A_0 is often fixed and calculated by the average across breathing cycles in previous studies (Lujan *et al* 2003, George *et al* 2005). In this study A_0 is considered an independent parameter in order to provide the model with extra variability, and its distribution is also considered to be normal fitted to the breathing data $\mathcal{N}(\mu_{A_0}, \sigma_{A_0})$.

2. AAE model. The AAE breathing models are based on artificial neural networks. First, an encoder computes a few *latent* parameters (a low dimensional embedding) that uniquely characterize each high-dimensional breathing signal. The number of low-dimensional latent variables is optimally configured. A decoder reconstructs the original breathing signal using the latent variables from the encoder. A training process using a large set of samples ensures that the decoder accurately reconstructs breathing signals and that each of the latent parameters is approximately distributed according to the Gaussian distribution $\mathcal{N}(0, 1)$. Using as few as 5 parameters, the AAE breathing models can generate patient-specific realistic breathing signals with high accuracy and variability in period and amplitude (Pastor-Serrano *et al* 2021), as opposed to the sinusoidal model always yielding perfect regular sinusoidal samples.

Once the models are obtained, artificial breathing signals are generated by sampling model parameters from their distributions, with each parameter combination resulting in a unique signal. A sinusoidal signal is thereby

obtained by sampling a period, amplitude and inhale position from $\mathcal{N}(\mu_T, \sigma_T)$, $\mathcal{N}(\mu_A, \sigma_A)$ and $\mathcal{N}(\mu_{A_0}, \sigma_{A_0})$. For the AAE breathing models, different signals are obtained by sampling the 5 latent parameters from a Gaussian distribution $\mathcal{N}(0, 1)$. For both models, the starting phase ψ—the starting point of delivery within the first breathing cycle—is sampled from a uniform distribution $\mathcal{U}(0, 2\pi)$.

Statistical evaluation of interplay

Testing robustness against interplay effects involves a statistical evaluation using a set of N different dose distributions and DVHs corresponding to N different breathing samples. A DVH(D): $D \in \mathbb{R} \rightarrow V \in \mathbb{R}$ is a function obtained for a given structure of interest that indicates the fraction of volume V that receives a dose greater than or equal to D. The quantity $V_f = \mathrm{DVH}(fD_p)$ indicates the fraction of the target volume that receives at least a certain percentage f of the prescribed dose D_p. Alternatively, the value $D_f = \mathrm{DVH}^{-1}(fV)$ represents the lowest dose received by at least a fraction f of the volume. We use typical values for these quantities in order to assess the adequacy of treatment plans, e.g. the D_{98} or dose that 98% of the volume receives. Additionally, we calculate the *homogeneity index* (*HI*), defined as $HI = (D_2 - D_{98})/D_p$, where D_{98} and D_2 are the dose received by the 98% and 2% of the volume. *HI*s quantify how uniformly the majority of the target volume is irradiated, with lower values indicating smaller differences between the dose delivered to different parts of the target. The $V_{107/95}$ indicates the fraction of the target volume that receives a dose outside the usually clinically accepted interval $(0.95D_p, 1.07D_p)$, and it is calculated as $V_{107/95} = V_{107} + (1 - V_{95})$.

Our interplay evaluation is based on comparing the distributions of D_2, D_{98}, *HI* and $V_{107/95}$, referred to as quantities of interest (Θ) in the remainder of this paper. We approximate these distributions using a collection of n_i percentiles Θ_i obtained from the N available computed Θ values, which are compiled into a *percentile vector* $\delta_\Theta = \{\Theta_i\}_{i=1}^{n_i}$. Subsequently, we assess the similarity between the results of different statistical analyses by comparing distributions of each quantity of interest Θ via the percentile vectors δ_Θ. If different statistical evaluations yield similar distributions, the analysis of interplay and conclusions drawn regarding the quality of the plan will approximately be the same.

Overview

After obtaining a treatment plan that satisfies the planning constraints and objectives, the interplay simulation proceeds as follows:

1. N different breathing signals are obtained either by randomly sampling the parameters of the breathing models, or by cropping 1000 random fragments from the original recorded signal, where the width of the slicing window is equal to the treatment length.
2. Using the N signals, treatment plan information and machine parameters, we calculate N interplay dose distributions. Each dose distribution results in a DVH from which the HI, $V_{107/95}$, D_2 and D_{98} are calculated. For each patient, the final robustness evaluation is based on first calculating N=1000 interplay dose distributions using fragments of the recorded signal, and subsequently analyzing the difference in δ_Θ between 4DCT and ITV plans.
3. We numerically compare distributions using the relative distribution error (RDE). For a quantity of interest and its corresponding vector δ_Θ, we quantify the RDE between two different distributions as

$$\mathrm{RDE}(\delta^N_{\Theta,1}, \delta^N_{\Theta,2}) = \frac{1}{n_i}\sum_{i=1}^{n_i} \frac{|\Theta_{i,1} - \Theta_{i,2}|}{\Theta_{ref}} \times 100, \tag{1}$$

   where $n_i = 3$ (median, 2 and 98 percentiles), and the reference value for the quantity of interest Θ_{ref} is used to compute the relative error and is obtained from a single interplay dose distribution corresponding to a sinusoidal with average period, amplitude and initial inhale position.

4. We perform a series of experiments to evaluate how using a limited number of samples, using artificial signals or ignoring breathing hysteresis compromises evaluation accuracy:
   (i) For a different number of breathing samples $N = \{20, 50, 100, 500, 1000\}$ of the recorded signal, we compute the distribution over the quantities of interest. We perform two independent statistical analyses for each number of breathing samples N using a different subset of N interplay dose distributions, resulting in two different vectors $\delta^N_{\Theta,1}$ and $\delta^N_{\Theta,2}$ that are compared via the RDE.

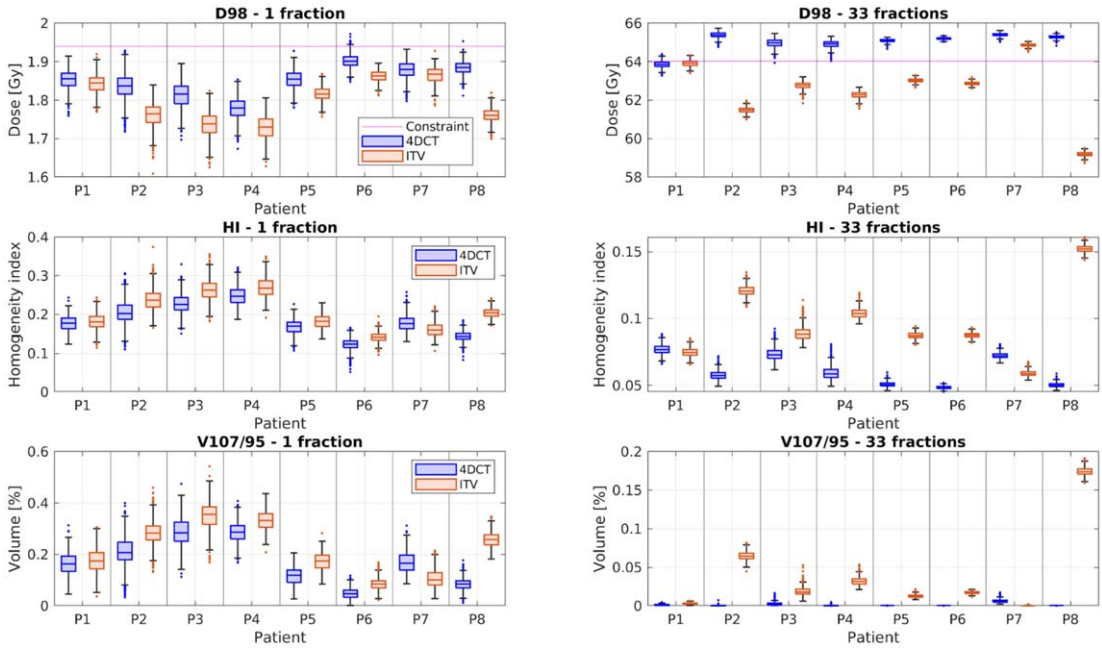


Figure 2. Interplay evaluation results. We display the distribution of 1000 different D_{98}, HI and $V_{107/95}$ CTV values for every patient and plan, and for (left) individual fractions and (right) fully fractionated treatments, using the real recorded signal. The pink line in the top row denotes the clinical near-minimum CTV dose constraint. For each box, the centered line represents the median, while the boundaries correspond to the upper and lower quartiles (25th and 75th percentiles), and the individual points outside the whiskers are outliers. Higher HI and $V_{107/95}$ correspond to more heterogeneous dose distributions with hot and cold spots.

(ii) To determine the effect of using artificial breathing signals from the sin or AAE models, we compute the RDE between their corresponding $\delta^{1000}_{\Theta,sin}$ or $\delta^{1000}_{\Theta,AAE}$ and the reference $\delta^{1000}_{\Theta,real}$ from the real recorded signals, where all the statistics are calculated using 1000 samples.

(iii) Finally, we assess the dosimetric impact of disregarding motion hysteresis by computing the RDE between the results of two different interplay evaluations with 1000 samples: one including 8 breathing phases, and the other only 5 phases identical during inhale and exhale.

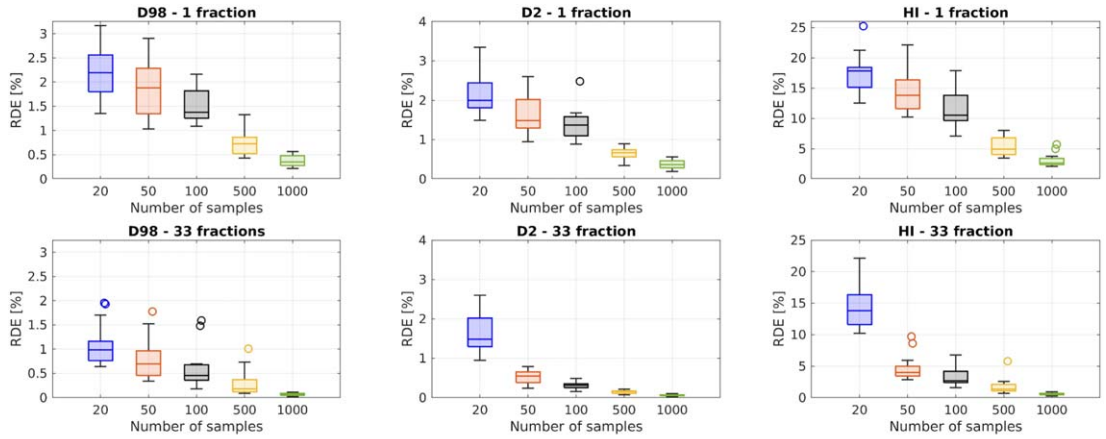
3. Results

Interplay robustness of 4DCT and ITV plans

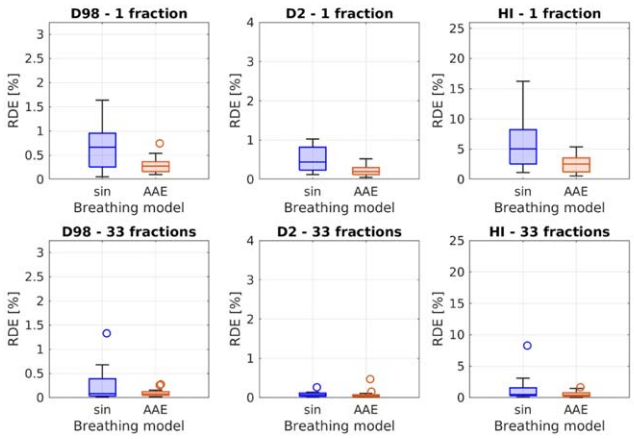
Reliable statistical analyses allow direct assessment of the robustness of treatment plans, as well as comparison between different planning approaches. To illustrate this, figure 2 shows the distribution of D_{98}, HI and $V_{107/95}$ corresponding to each plan and patient combination, for both single fractions and a fully fractionated treatment. As seen in the top row, the 4DCT plans result in higher D_{98} values regardless of fractionation, tumor size and breathing amplitude, while ITV plans systematically fail to meet the clinical constraints. Likewise, the HI and $V_{107/95}$ (middle and bottom rows) are consistently lower in 4DCT treatments, indicating that the delivered dose distribution is more homogeneous and the target receives a dose within the clinically acceptable limits in most of the scenarios.

Influence of sample size, breathing models and hysteresis.

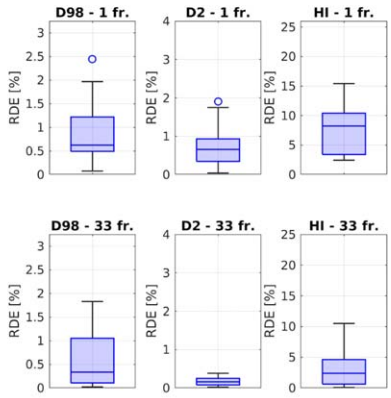
A relevant question is how many different interplay dose distributions are necessary in order to perform a statistical analysis that yields reliable results. For this reason, independent statistical analyses are performed using different sub-sample sizes selected according to published results (Seco *et al* 2009, Engwall *et al* 2018a, 2018b, Jakobi *et al* 2018). Figure 3(a) shows a reduction in RDE as more breathing samples are used to calculate the statistics, confirming that the distributions gradually converge. The RDE illustrates how much the results from two statistical analysis could vary for a given sample size simply due to chance, being higher for single fraction analyses using <100 samples. Figure 3(b) shows a comparison of the error introduced by using artificial signals from the sin and AAE models instead of the recorded signals from the patient, showing that for single fractions AAE model slightly outperforms the sin model, but the differences between models mostly fade in fully fractionated treatments. Figure 3(c) shows the effect of disregarding breathing hysteresis. While the



(a) Impact of using different sample sizes to calculate interplay statistics during evaluation.



(b) Dosimetric effect of using breathing signals from a model instead of the real recorded signals.



(c) Impact of ignoring breathing hysteresis during the evaluation.

Figure 3. Effect of the evaluation parameters on the interplay statistics. The reported RDE represents the difference between two different distributions of a quantity of interest, in this case the D_{98}, D_2 and *HI*, and can be used to determine whether two independent interplay evaluations yield the same results. For each box showing RDE values across patients and planning approaches, we evaluate the dosimetric impact of varying one of the following evaluation parameters, while keeping the rest fixed: (a) the number of samples used to compute the statistics, (b) the breathing signal model, and (c) the absence of respiratory motion hysteresis, with identical inhale and exhale. Each variation results in an independent distribution, which is compared to either (a) a duplicate distribution obtained using the same settings, or (b), (c) a reference distribution obtained from a statistical analysis using 1000 samples from the recorded signal and considering breathing hysteresis. Each box contains the median in the center and the upper and lower quartiles (25th and 75th percentiles) as box boundaries, with outliers represented as individual points outside the whiskers.

model signals and symmetrical respiratory motion have a lower dosimetric impact than using few samples in the evaluation, the errors are high for single fractions and overall non-negligible, especially for the *HI*.

Interplay dose dependence on breathing parameters.

In order to investigate the relationship between small changes in the breathing parameters and interplay doses, figure 4 shows the dose and D_{98} for different amplitudes, periods and starting phases of a sinusoidal breathing signal. Each of the parameters is varied independently, one at a time, leaving the rest fixed. Amplitude changes have a lower and less fluctuating effect on the dose compared to changes in period or starting phase. The latter affect the time structure of treatment delivery, and as a result, small variations can effectively shift the breathing phases in which subsequent spots are delivered, with a great local impact on voxel doses. On the other hand, changes in amplitude are responsible for shifting only few spots to neighboring phases, hence inducing smaller changes in the delivered dose. Repainting contributes to better target coverage and reducing the magnitude of interplay effects, as indicated by the lower spread of voxel doses around the target 2 Gy fraction dose, and the higher D_{98} values. For fractions delivered without repainting, period changes can result in up to 50% variations over the target dose and 4 Gy differences in D_{98}, as seen in the top row of figure 4.

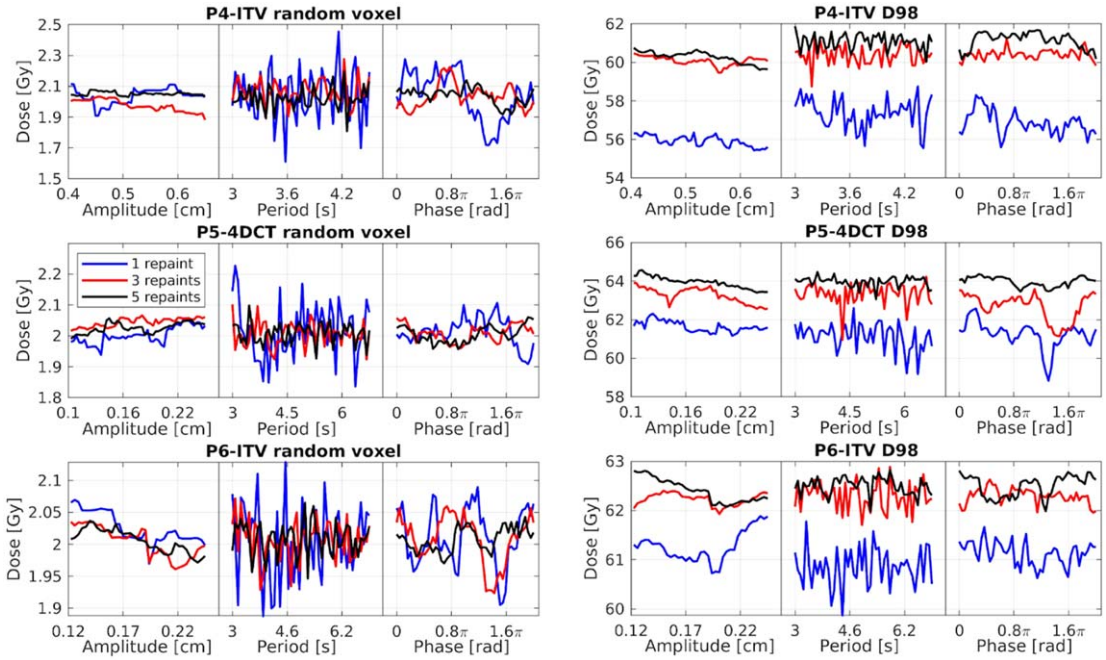


Figure 4. Effect of breathing parameters on the final dose distribution. (Left) Fraction dose in a random CTV voxel and dependence on the parameters of a sinusoidal breathing signal, for a different number of repaintings and 3 patients (from top to bottom, patient 4, patient 5 and patient 6). Blue lines correspond to dose distributions without repainting, whereas red and black lines indicate 3 and 5 repaints. Repainting smooths out interplay effects in the local fraction doses and reduces the fluctuations around the 2 Gy target dose. As a result Similar results are obtained for other randomly selected voxels. (Right) D_{98} dependence on small breathing variations. Repainted dose distributions fluctuate less and result in better target coverage, as indicated by the higher D_{98} values.

4. Discussion

Robustness evaluation of 4DCT and ITV plans

Our results indicate that 4DCT plans outperform ITV plans in terms of dose coverage and homogeneity, regardless tumor size and breathing amplitude. Using a fully fractionated robust 4DCT treatment planning approach with the exhale, inhale and mid-ventilation phases may be sufficient to compensate the detrimental effect of breathing motion, as indicated by the high D_{98} values and lower HI and $V_{107/95}$ shown in figure 2. Contrariwise, robust ITV plans seem to fail to meet the required dose constraints in IMPT lung cancer treatments, and may require the use of additional margins or motion mitigation techniques, or increased robustness settings. Finally, as seen in the left plots of figure 2 by the lower D_{98} and higher HI and $V_{107/95}$ in single fraction doses, interplay effects seem to be aggravated by larger amplitudes and tumor sizes (P3 and P4).

Accuracy of the interplay evaluation

Among all possible simplifications (i.e. using few breathing samples or ignoring hysteresis), the error of using artificial signals seems to be the lowest, where the AAE breathing model clearly outperforms the sin model at a considerably higher computational cost and more patient-specific data. In the light of our results, a simple sinusoidal model may be sufficiently accurate in fully fractionated treatments as long as the parameter distribution is patient specific. Disregarding hysteresis, however, introduces errors that can be as high as 2.5% of the D_{98} of the delivered dose in some cases, even when considering the smoothing effect of fractionation (figure 3(c)).

Using a few realizations (<100) of interplay dose distributions in order to evaluate interplay effects lacks statistical power. Our results from lung cancer patients show that at least 500 different interplay dose distributions are needed to achieve the same level of error as the one introduced by other simplifications such as using sinusoidal breathing or no hysteresis, also for fractionated delivery. Only for >500 samples are used the differences are generally below 1% of the reference dose and 5% of the HI values, which can be limiting with computationally expensive interplay dose calculation models. Most of the previous studies are short on samples: ranging from 300 different simulated treatments (Seco *et al* 2009) to as few as 10 samples (Engwall *et al* 2018a), 20 samples (Engwall *et al* 2018b) or 64 samples (Jakobi *et al* 2018). Other published works do not explicitly reference this number but use few realizations with different starting phases (Kardar *et al* 2014, Li *et al* 2014), or are based on 30 dose distributions weighted by their probability of occurrence (Kraus *et al* 2011).

The combined smoothing benefits of repainting and fractionation in lung cancer treatments has been previously investigated (Seco *et al* 2009, Li *et al* 2014, Engwall *et al* 2018a) and is further exemplified in figure 4. We therefore assume that the worst-case scenario in the interplay dose degradation occurs for single fraction dose distributions with no motion mitigation, which explains the fact that the errors in the statistical evaluation diminish as fractionation increases. As a result, the relative errors between distributions may be minimal if repainting or other mitigation technique are applied, requiring fewer samples to obtain reliable results and thus compensating for the longer calculation times needed to simulate repainting.

We focus this research on IMPT lung cancer patients, that represent a worst-case scenario for breathing motion. Other treatment modalities such as SBRT or hypo-fractionated IMPT treatments deliver the dose more intensely using less fractions. The considerably higher dose per fraction could exacerbate interplay effects (and in particular may cause bigger inhomogeneities in the dose), especially in terms of biological dose. For such cases, evaluating the dose degradation due to motion using only few samples could lead to even larger inaccuracies.

Dose dependence on breathing parameters

The results in figure 4 demonstrate the beneficial effect of repainting in both smoothing out great local dose variations and improving target coverage, as seen in the reduced fluctuations around the 2 Gy target dose that translate into higher D_{98} values. However, rescanning alone does not fully mitigate interplay effects, in concordance with previous results (Zhang *et al* 2018), resulting in local doses that may vary up to 10% of the target dose and D_{98} values that are always below the constraint. Delivery without repainting results in dose fluctuations amounting up to 50% of the target fraction dose. We deduce that this effect is caused by the fact that small period and starting phase changes can simultaneously shift a significant number of subsequent spots, the effect being more dramatic for the parts of the tumor that receive dose only from few individual pencil beams, or spots delivered later within a fraction. Our results are consistent with previous findings for IMRT dynamic delivery (Kissick *et al* 2005) that demonstrate the detrimental effect of intra-fraction random changes of the breathing parameters. We further hypothesize that these results are independent of the 4DCT resolution: adding more 4DCT phases during evaluation results in some spots shifting to consequent phases with similar anatomy, and thus the effect may not as dramatic as with period or phase changes, where small variations may cause the delivery of a later spot to shift from full inhale to exhale.

The degrading effects of time changes can also impact currently applied clinical protocols. Most of the treatment centers establish their criteria for interplay mitigation in terms of breathing amplitude (e.g. no mitigation is considered if the breathing amplitude for a given patient is lower than 5 mm). We show that not only does period influence the fluctuating behavior but it also highly affects the degree of degradation of the dose. Thus, more research is needed to determine whether making planning or clinical decisions purely based on amplitude criteria suffices, and whether strategies that weigh both period and amplitude changes offer additional benefits.

Limitations

The most limiting design choice is the use of a single 4DCT, under the assumption that it captures the variations in patient anatomy from full inhale (maximum amplitude) to full exhale and breathing hysteresis, as well as the mismatch between 4DCT and the signal motion surrogate. While we make this assumption to speed-up and simplify the interplay dose distribution calculation, some irregularities may not be captured in the 4DCT, for which a bio-mechanical model could be used to simulate hiccups or coughs as in Boye *et al* (2013). Similarly, the temporal resolution of the 4DCT is significantly lower than that of the spot delivery. Although we do not expect the effects of a coarser resolution to be as significant as disregarding hysteresis, the most detailed interplay simulations should be based on variable time dependent 4DCT data with finer temporal resolution. Finally, the accuracy and calculation times of the presented interplay dose calculation method ultimately depend on that of the dose engine and the registration algorithm. Traditional (usually slow) image registration methods have been been recently outperformed by data-driven approaches (Balakrishnan *et al* 2018, 2019, Dalca *et al* 2018, 2019). Similarly, recent deep learning based dose engines (Wu *et al* 2021, Pastor-Serrano and Perk 2021) have been shown to overcome the speed limitations of Monte Carlo methods, while offering better performance than the pencil beam algorithms commonly used in the clinics.

5. Conclusions

We present a practical method to simulate dose delivery under motion interplay effects and assess treatment robustness based on hundreds of (sampled) breathing signals. Our statistical evaluation shows that ITV plans systematically fall behind their computationally more expensive 4DCT robust counterpart, regardless of tumor size and breathing amplitude. After analyzing the error introduced by simplifications such as neglecting motion

hysteresis or using few interplay scenarios and sinusoidal breathing signals, we conclude that the statistical analysis of fully fractionated treatments requires at least 500 different dose distributions corresponding to 500 different samples of regular sinusoidal breathing (based on patient-specific parameter distributions) with hysteresis to yield acceptable precision. We complete this study by demonstrating that small breathing period variations have a highly non-linear effect on local dose deposition and can cause local doses to fluctuate up to 50% of the target fraction dose.

Acknowledgments

This work is funded by KWF Kanker Bestrijding, and part of the KWF research project PAREL with grant number 11 711. The authors are particularly grateful to Dr Abdul Wahab Sharfo (Erasmus MC, Rotterdam) for the discussions and providing the necessary data, and Dr. David Boersma for the discussion and revision of the manuscript. Zoltán Perkó would like to thank the support of the NWO VENI grant ALLEGRO (016. Veni.198.055) during the time of this study.

CRediT authorship contribution statement

Oscar Pastor-Serrano: conceptualization, Data Curation, Methodology, Investigation, Formal Analysis, Visualization, Software, Writing—original draft.

Steven Habraken: resources, Formal Analysis, Data curation, Writing—review and editing.

Danny Lathouwers: supervision, Writing—review and editing.

Mischa Hoogeman: conceptualization, Formal Analysis, Project administration, Funding acquisition, Writing—review and editing.

Dennis Schaart: conceptualization, Supervision, Writing—review and editing.

Zoltán Perkó: conceptualization, Supervision, Methodology, Formal Analysis, Funding acquisition, Project administration, Writing—review and editing.

ORCID iDs

Oscar Pastor-Serrano https://orcid.org/0000-0002-2328-1429
Dennis Schaart https://orcid.org/0000-0002-3199-5608
Zoltán Perkó https://orcid.org/0000-0002-0975-4226

References

Bert C and Durante M 2011 Motion in radiotherapy: particle therapy *Phys. Med. Biol.* **56** R113–44

Bert C, Gemmel A, Saito N and Rietzel E 2009 Gated irradiation with scanned particle beams *Int. J. Radiat. Oncol. Biol. Phys.* **73** 1270–5

Bert C, Grözinger S O and Rietzel E 2008 Quantification of interplay effects of scanned particle beams and moving targets *Phys. Med. Biol.* **53** 2253–65

Bert C, Saito N, Schmidt A, Chaudhri N, Schardt D and Rietzel E 2007 Target motion tracking with a scanned particle beam *Med. Phys.* **34** 4768–71

Bernatowicz K, Zhang Y, Perrin R, Weber D C and Lomax A J 2017 Advanced treatment planning using direct 4D optimisation for pencil-beam scanned particle therapy *Phys. Med. Biol.* **62** 6595–609

Boye D, Lomax T and Knopf A 2013 Mapping motion from 4D-MRI to 3D-CT for use in 4D dose calculations: a technical feasibility study *Med. Phys.* **40** 061702

Bortfeld T, Jokivarsi K, Goitein M, Kung J and Jiang S B 2002 Effects of intra-fraction motion on IMRT dose delivery: statistical analysis and simulation *Phys. Med. Biol.* **47** 2203-20

Breedveld S, Storchi P R, Voet P W and Heijmen B J 2012 iCycle: integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans *Med. Phys.* **39** 951–63

Balakrishnan G, Zhao A, Sabuncu M R, Guttag J and Dalca A V 2019 Voxelmorph: a learning framework for deformable medical image registration *IEEE Trans. Med. Imaging* **38** 1788–800

Balakrishnan G, Zhao A, Sabuncu M R, Guttag J and Dalca A V 2018 An unsupervised learning model for deformable medical image registration *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 9252–60

Boda-Heggemann J *et al* 2016 Deep inspiration breath hold-based radiation therapy: a clinical review *Int. J. Radiat. Oncol. Biol. Phys.* **94** 478–92

Court L E, Wagar M, Ionascu D, Berbeco R and Chin L 2008 Management of the interplay effect when using dynamic MLC sequences to treat moving targets *Med. Phys.* **35** 1926–31

Court L, Wagar M, Berbeco R, Reisner A, Winey B, Schofield D, Ionascu D, Popple R and Lingos T 2010 Evaluation of the interplay effect when using RapidArc to treat targets moving in the craniocaudal or right-left direction *Med. Phys.* **37** 4–11

Coste-Manire E, Olender D, Kilby W and Schulz R A 2005 Robotic whole body stereotactic radiosurgery: clinical advantages of the Cyberknife integrated system *Int. J. Med. Robot. Comput. Assist. Surg.: mrcAS* **1** 28–39

Dalca A V, Balakrishnan G, Guttag J and Sabuncu M R 2018 Unsupervised learning for fast probabilistic diffeomorphic registration *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 729–38

Dalca A V, Balakrishnan G, Guttag J and Sabuncu M R 2019 Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces *Med. Image Anal.* **57** 226–36

den Boer E, Wulff J, MÃder U, Engwall E, BÃumer C, Perko Z and Timmermann B 2021 Investigating interplay effects in pencil beam scanning proton therapy with a 4D XCAT phantom within the RayStation treatment planning system *Med. Phys.* **48** 1448–55

Evans P M, Coolens C and Nioutsikou E 2005 Effects of averaging over motion and the resulting systematic errors in radiation therapy *Phys. Med. Biol.* **51** N1

Engwall E, Glimelius L and Hynning E 2018a Effectiveness of different rescanning techniques for scanned proton radiotherapy in lung cancer patients *Phys. Med. Biol.* **63** 095006

Eley J G, Newhauser W D, LÃchtenborg R, Graeff C and Bert C 2014 4D optimization of scanned ion beam tracking therapy for moving tumors *Phys. Med. Biol.* **59** 3431–52

Engwall E, Fredriksson A and Glimelius L 2018b 4D robust optimization including uncertainties in time structures can reduce the interplay effect in proton pencil beam scanning radiation therapy *Med. Phys.* **45** 4020–9

Emert F, Missimer J, Eichenberger P A, Walser M, GmÃr C, Lomax A J and Spengler C M 2021 Enhanced deep-inspiration breath hold superior to high-frequency percussive ventilation for respiratory motion mitigation: a physiology-driven, MRI-guided assessment toward optimized lung cancer treatment with proton therapy *Front. Oncol.* **11** 621350

Engelsman M, Rietzel E and Kooy H M 2006 Four-dimensional proton treatment planning for lung tumors *Int. J. Radiat. Oncol. Biol. Phys.* **64** 1589–95

George R, Vedam S S, Chung T D, Ramakrishnan V and Keall P J 2005 The application of the sinusoidal model to lung cancer patient respiratory motion *Med. Phys.* **32** 2850–61

Hoogeman M, PrÃvost J B, Nuyttens J, PÃll J, Levendag P and Heijmen B 2009 Clinical accuracy of the respiratory tumor tracking system of the cyberknife: assessment by analysis of log files *Int. J. Radiat. Oncol. Biol. Phys.* **74** 297–303

Heath E, Unkelbach J and Oelfke U 2009 Incorporating uncertainties in respiratory motion into 4D treatment plan optimization *Med. Phys.* **36** 3059–71

Jiang S B, Pope C, Al Jarrah K M, Kung J H, Bortfeld T and Chen G T 2003 An experimental investigation on intra-fractional organ motion effects in lung IMRT treatments *Phys. Med. Biol.* **48** 1773

Jakobi A, Perrin R, Knopf A and Richter C 2018 Feasibility of proton pencil beam scanning treatment of free-breathing lung cancer patients *Acta Oncol.* **57** 203–10

Kissick M W, Boswell S A, Jeraj R and Mackie T R 2005 Confirmation, refinement, and extension of a study in intrafraction motion interplay with sliding jaw motion *Med. Phys.* **32** 2346–50

Kraus K M, Heath E and Oelfke U 2011 Dosimetric consequences of tumour motion due to respiration for a scanned proton beam *Phys. Med. Biol.* **56** 6563–81

Krieger M, Giger A, Salomir R, Bieri O, Celicanin Z, Cattin P C, Lomax T, Weber D and Zhang Y 2020 Impact of internal target volume definition for pencil beam scanned proton treatment planning in the presence of respiratory motion variability for lung cancer: a proof of concept *Radiother. Oncol.* **145** 154–61

Kardar L *et al* 2014 Evaluation and mitigation of the interplay effects of intensity modulated proton therapy for lung cancer in a clinical setting *Pract. Radiat. Oncol.* **4** e259–68

Li Y *et al* 2014 On the interplay effects with proton scanning beams in stage III lung cancer *Med. Phys.* **41** 021721

Lujan A E, Balter J M and Ten Haken R K 2003 A method for incorporating organ motion due to breathing into 3D dose calculations in the liver: sensitivity to variations in motion *Med. Phys.* **30** 2643–9

Lambert J, Suchowerska N, McKenzie D R and Jackson M 2005 Intrafractional motion during proton beam scanning *Phys. Med. Biol.* **50** 4853–62

Ohara K, Okumura T, Akisada M, Inada T, Mori T, Yokota H and Calaguas M J 1989 Irradiation synchronized with respiration gate *Int. J. Radiat. Oncol. Biol. Phys.* **17** 853–7

Pastor-Serrano O and Perkó Z 2021 Learning the Physics of Particle Transport via TransformersarXiv:2109.03951 [cs.LG]

Pflugfelder D, Wilkens J J and Oelfke U 2008 Worst case optimization: a method to account for uncertainties in the optimization of intensity modulated proton therapy *Phys. Med. Biol.* **53** 1689–700

Phillips M H, Pedroni E, Blattmann H, Boehringer T, Coray A and Scheib S 1992 Effects of respiratory motion on dose uniformity with a charged particle scanning method *Phys. Med. Biol.* **37** 223–34

Pguret N *et al* 2016 Apnea-like suppression of respiratory motion: first evaluation in radiotherapy *Radiother. Oncol.* **118** 220–6

Pastor-Serrano O, Lathouwers D and Perkó Z 2021 A semi-supervised autoencoder framework for joint generation and classification of breathing *Comput. Methods Programs Biomed.* **209** 106312

Seco J, Robertson D, Trofimov A and Paganetti H 2009 Breathing interplay effects during proton beam scanning: simulation and statistical analysis *Phys. Med. Biol.* **54** N283–94

Seco J, Sharp G C, Turcotte J, Gierga D, Bortfeld T and Paganetti H 2007 Effects of organ motion on IMRT treatments with segments of few monitor units *Med. Phys.* **34** 923–34

Shih H A, Jiang S B, Aljarrah K M, Doppke K P and Choi N C 2004 Internal target volume determined with expansion margins beyond composite gross tumor volume in three-dimensional conformal radiotherapy for lung cancer *Int. J. Radiat. Oncol. Biol. Phys.* **60** 613–22

van de Water S, Kraan A C, Breedveld S, Schillemans W, Teguh D N, Kooy H M, Madden T M, Heijmen B J and Hoogeman M S 2013 Improved efficiency of multi-criteria IMPT treatment planning using iterative resampling of randomly placed pencil beams *Phys. Med. Biol.* **58** 6969–83

Wu C, Nguyen D, Xing Y, Montero A B, Schuemann J, Shang H and Jiang S 2021 Improving proton dose calculation accuracy by using deep learning *Mach. Learn.: Sci. Technol.* **2** 015017

Yu J, Park S S, Herman M G, Langen K, Mehta M and Feigenberg S J 2017 Free breathing versus breath-hold scanning beam proton therapy and cardiac sparing in breast cancer *Int. J. Part. Ther.* **3** 407–13

Zhang Y, Knopf A, Tanner C and Lomax A J 2014 Online image guided tumour tracking with scanned proton beams: a comprehensive simulation study *Phys. Med. Biol.* **59** 7793–817

Zhang Y, Huth I, Wegner M, Weber D C and Lomax A J 2016 An evaluation of rescanning technique for liver tumour treatments using a commercial PBS proton therapy system *Radiother. Oncol.* **121** 281–7

Zhang Y, Huth I, Weber D C and Lomax A J 2018 A statistical comparison of motion mitigation performances and robustness of various pencil beam scanned proton systems for liver tumour treatments *Radiother. Oncol.* **128** 182–8