Clinical 3D/4D cumulative proton dose assessment methods for thoracic tumours with large motion

Sabine Visser*, Erik W. Korevaar, Christina T. Muijs, Robin Wijsman, Johannes A. Langendijk, Pietro Pisciotta, Gabriel Gutteres Marmitt, Cássia O. Ribeiro, Stefan Both

Department of Radiation Oncology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands

Abstract

Purpose: Despite the anticipated clinical benefits of intensity-modulated proton therapy (IMPT), plan robustness may be compromised due to its sensitivity to patient treatment uncertainties, especially for tumours with large motion. In this study, we investigated treatment course-wise plan robustness for intra-thoracic tumours with large motion comparing a 4D pre-clinical evaluation method (4DREM) to our clinical 3D/4D dose reconstruction and accumulation methods.

Materials and methods: Twenty patients with large target motion (>10 mm) were treated with five times layered rescanned IMPT. The 3D-robust optimised plans were generated on the averaged planning 4DCT. Using multiple 4DCTs, treatment plan robustness was assessed on a weekly and treatment course-wise basis through the 3D robustness evaluation method (3DREM, based on averaged 4DCTs), the 4D robustness evaluation method (4DREM, including the time structure of treatment delivery and 4DCT phases) and 4D dose reconstruction and accumulation (4DREAL, based on fraction-wise information).

Results: Baseline target motion for all patients ranged from 11–17 mm. For the offline adapted course-wise dose assessment, adequate target dose coverage was found for all patients. The target volume receiving 95% of the prescription dose was consistent between methods with 16/20 patients showing differences < 1%. 4DREAL showed the highest target coverage (99.8 ± 0.6%, p < 0.001), while no differences were observed between 3DREM and 4DREM (99.3 ± 1.3% and 99.4 ± 1.1%, respectively).

Conclusion: Our results show that intra-thoracic tumours can be adequately treated with IMPT in free breathing for target motion amplitudes up to 17 mm employing any of the accumulation methods.

Keywords: Lung cancer, Oesophageal cancer, Large motion, Intensity-modulated proton therapy, 4D dose assessment, Dose accumulation

Tumours located in the thoracic region are challenging for radiotherapy due to their proximity to critical organs-at-risk (OARs). For lung and oesophageal cancer, the healthy lung and oesophageal tissue, the heart and the spinal cord are the most relevant structures to be dosimetrically spared. Compared to photon radiotherapy, large OAR dose reductions may be attained with intensity-modulated proton therapy (IMPT) due to the physical properties of protons (allowing for a sharp distal dose fall-off with virtually no exit dose) [1,2]. OAR dose reduction is important as numerous studies have demonstrated the relationship between OARs doses and observed toxicities after radiotherapy for intra-thoracic tumours, particularly concerning pulmonary complications [3,4].

Besides the complex anatomy of the thorax, radiotherapy for this region may be affected by respiratory motion and anatomical changes. Especially for proton therapy, robustness may be compromised due to its sensitivity to various sources of range errors [5]. Robust optimisation takes multiple range and setup scenarios into account to ensure plan robustness and has been implemented by several centres for the treatment of intra-thoracic tumours with small to moderate breathing amplitudes using the internal target volume (ITV) concept [2,6,7]. However, the continuous breathing movement of the target volume during pencil beam scanning may create over- and/or underdosage of the target volume, generally referred to as interplay effects. In small to moderate movers, these effects seem limited when fractionation is considered [7–11]. Interfractional anatomical changes might have more considerable influence on the dose distributions [12,13]. Currently, a limited number of studies combine all possible disturbances to confirm target dose coverage robustness during the treatment course. In our clinic, we have developed a comprehensive 4D
robustness evaluation method (4DREM) to assess the treatment course-wise robustness of IMPT plans before its clinical deployment [14]. This method encompasses all possible uncertainties, such as setup and range errors, machine errors, patient anatomy changes, breathing motion and interplay effects. For intra-thoracic tumours with moderate motion, we found that our five times layer rescanned 3D-optimised IMPT plans ensure adequate target coverage, giving us the confidence to start treating these patients [7]. However, 4DREM is computationally and time intensive, making it not suitable for routine clinical use, especially when we aim to reconstruct every patients’ treatment fraction in the clinic. Therefore, in clinical practice, we currently employ an accelerated and simplified fraction-wise accumulated 4D dose reconstruction (4DREAL) combined with a weekly average CT-based robustness evaluation method (3DREM) [13,15,16]. This combined 3DREM and 4DREAL evaluation is only performed for patients with large target motion (clinical target volume [CTV] motion ≥ 10 mm).

In this work, we show the first robustness results for thymoma, lung and oesophageal cancer patients with large motion amplitudes treated at our proton centre with IMPT, in terms of target coverage and OARs doses. For this high-risk patient group, we compare our clinical evaluation methods (4DREAL and 3DREM) to our pre-clinical evaluation method (4DREM) on a weekly and course-wise basis to determine the adequacy of the clinical evaluation methods and investigate the impact of considered errors.

Methods and materials

Patient and plan data

For this study, we selected the first twenty intra-thoracic indications with large CTV motion (10–20 mm) who were treated in free breathing setting at our proton centre. These included thymoma, lung and oesophageal cancer patients, all selected for proton therapy using the model-based approach [17]. The patients were enrolled in our standardised follow-up programme, approved by the medical ethics committee (METc2014.379), for which all (surviving) patients provided informed consent. All patients underwent a 4DCT for planning, and weekly 4DCTs for verification. Each 4DCT was reconstructed in ten breathing phases using phase-wise minimum dose distribution (Vw min) reconstructed from all 28 scenarios. The resulting voxel-wise maximum dose distribution on the ITV had to be at least 94% of the prescription dose on the voxel-wise minimum dose distribution (Vw min) reconstructed from all 28 scenarios. The resulting voxel-wise maximum dose distribution was inspected for hotspots.

Dose verification and plan adaptation

During the course of the proton treatment, CBCTs were acquired daily for position verification purposes. The weekly verification 4DCT was registered to the daily CBCTs for that week and the fused images were visually checked by a medical physicist to assess the representativity of the verification 4DCT, which resulted in exclusion of two 4DCTs due to different positioning (patient 5 – week 4 and patient 10 - week 2). A 3D robustness evaluation was performed on the verification 4DCT, and the resulting dose distribution was checked by a radiation oncologist, and used as trigger for clinical plan adaptation. For the lung cancer patients included in this study, only one plan adaptation was performed for patient 2 due to hotspots appearing outside of the target volume. For the oesophageal cancer patients, five plan adaptations were performed in total for four patients. For patient 12, replanning was necessary since the patient had to change to an arms-down position after the first week of treatment. For the remaining planned oesophageal cancer patients, target dose coverage was affected due to diaphragm displacements and/or target deformations. In the following evaluations, we assessed the patients’ complete clinical treatment course including adaptation considered from the fraction it was initiated in the clinic.

Weekly and full course dose evaluation

The dose was reconstructed every week and for the full treatment course using 3DREM, 4DREM and 4DREAL. The weekly evaluation uses one verification 4DCT of that specific week, and the course-wise evaluation considers all available 4DCTs equally. All three assessments are based on dose accumulation on the baseline CT (the initial planning CT) after dose warping using deformable image registration (DIR) and multiple 4DCTs. The DIRs used for dose warping between images were visualised and checked by a medical physicist. The DIR method used throughout this study was the ANACONDA algorithm available in our TPS [18]. The details of each evaluation are summarised in Table 1 and are described in more detail in the following sections.

3DREM

In the 3DREM evaluation, 28 scenarios are created on the verification averaged 4DCT(s) including a ±3% range uncertainty and again fourteen setup error scenarios in x, y and z directions with shifts of 2 mm in total [16,19,20]. For each of the 28 scenarios, the dose is warped from the respective weekly verification averaged 4DCT to the baseline averaged 4DCT. Here, a Vw min is reconstructed from the 28 scenarios.

4DREM

In the 4DREM evaluation, fourteen scenarios are calculated with each scenario encompassing eight fractions in total [7,14]. In each scenario, a systematic range error (0% or ± 3%) is considered. Additionally, a systematic (fourteen shifts similarly as 3DREM) and a random setup error (per fraction) are included, providing a total magnitude of 2 mm following the van Herk recipe [7,21]. The random part of the setup error is sampled from a normal distribution. To split the treatment plan over the ten 4DCT phases, machine log files (resulting from a single dry run of the plan) and a 4.5 s constant breathing cycle are used. Subsequently, the doses on the
4DCT phases are warped to the end-of-exhale phase of the corresponding 4DCT according to the deformation vector fields of the DIR. The fraction dose is accumulated here, before being warped to the end-of-exhale phase of the baseline 4DCT and here, the fraction doses are accumulated. In the end, a $V_{W_{\text{min}}}$ is reconstructed from the fourteen scenario doses.

**4DREAL**

In the 4DREAL evaluation, the delivery log files and the patient breathing pattern records, acquired through the Anzai belt system (Anzai Medical, Tokyo, Japan), are collected for every treatment fraction. The method uses the patient treatment delivery information to split the plan over the ten breathing phases of the verification 4DCT. Then, to reconstruct the daily treatment fraction dose on the end-of-exhale phase, dose warping and accumulation is performed identically as for 4DREM, using the same deformation vector fields from DIR. The daily fraction doses are then accumulated on the end-of-exhale phase of the baseline 4DCT [13,15].

**Evaluated metrics**

Weekly and course-wise target coverage was evaluated in terms of the volume of the target that receives 95% of the prescribed dose ($V_{95}$). For evaluation in 4DREAL and 4DREM, the CTV on the end-of-exhale phase of the baseline 4DCT was used as the target volume. Target evaluation in 3DREM was performed on the ITV on the averaged baseline 4DCT. In 3DREM and 4DREM, the $V_{95}$ was determined from the $V_{W_{\text{min}}}$ reconstructed from the scenario doses. Additionally, OARs doses and target homogeneity ($D_{2}-D_{98}$) were evaluated course-wise, based on all scenarios and reported using the mean and standard deviation (SD). Investigated OAR dose metrics were the mean lung dose (MLD, lungs minus gross tumour volume) and the mean heart dose (MHD).

**Results**

The patient population included in this study consisted of thymoma, lung and oesophageal cancer patients treated at the proton therapy centre (Table 2). For all patients, large motion (11–17 mm) was observed in the point maximum motion analysis of the CTV. Mean CTV motion based on the baseline 4DCT ranged from 1.5–10.9 mm and was not always consistent with the verification 4DCTs. For patient 19 for instance, the mean motion on the verification 4DCTs was 19.7 ± 4.6 mm, which is considerably higher than the mean CTV motion on the baseline 4DCT (7.0 mm).

The 4DREAL showed the highest target coverage for all patients (99.8 ± 0.6%, $p < 0.001$). Target homogeneity in 4DREAL was also superior for all but patient 16, comparing its value to the mean of the 3DREM and 4DREM scenarios. For patients with underdosage of the target, the homogeneity index was higher and the large SD indicated more inconsistencies between scenarios. Differences in target coverage were less than 1% in 19/20 patients between 3DREM and 4DREM. There was no trend observed between target coverage in 3DREM and 4DREM (99.3 ± 1.3% and 99.4 ± 1.1% respectively, $p = 0.268$), and for the two mentioned cases with a lower $V_{95}$ (#17 and #19), differences between meth-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Overview of the three cumulative proton dose assessment methods and their specifications. The number of scenarios is given in brackets the number of systematic setup errors times the number of range errors considered for each setup error.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>3DREM</strong></td>
</tr>
<tr>
<td><strong>CT image used for dose evaluation</strong></td>
<td>Averaged 4DCT</td>
</tr>
<tr>
<td><strong>Range error</strong></td>
<td>-3 and +3 %</td>
</tr>
<tr>
<td><strong>Setup error</strong></td>
<td>2 mm (systematic)</td>
</tr>
<tr>
<td><strong>Number of scenarios</strong></td>
<td>28 (14x2)</td>
</tr>
<tr>
<td><strong>Fractions</strong></td>
<td># of verification CTs</td>
</tr>
<tr>
<td><strong>Log file based</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>Breathing signal</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>Interplay</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>Machine errors</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>Time estimation</strong></td>
<td>180 minutes</td>
</tr>
</tbody>
</table>

*Abbreviations: 3DREM = 3D robustness evaluation method, 4DREM = 4D robustness evaluation method, 4DREAL = 4D dose reconstruction and accumulation. *Based solely on the calculation/warping/accumulation process for a two-field treatment plan with five repeated CTs (therefore excluding plan delivery, data collection/preparation and subplan creation).
ods were small and underdosage appeared in the same regions of the target (Fig. 2A). Although the $V_{95}$ was less than 98% of the prescribed dose for these two patients, underdosage appeared in the elective region of the target volume, which was deemed not clinically relevant by the treating physician. Based on the weekly evaluation, clinical plan adaptations were performed once for patient 17 and twice for patient 19, due to a different diaphragm position, which improved target coverage in the following week in five out of nine evaluations (Fig. 2B). Furthermore, motion extracted from the verification 4DCTs was the largest for these patients. Target coverage was further investigated based on weekly verification and presented for all patients in the supplementary materials (Supp. C). Also here, target coverage was higher in the 4DREAL evaluation, and no trends were observed between 3DREM and 4DREM. For 9/10 lung cancer patients, the difference between 3DREM and 4DREM was below 1%. Larger differences were observed for the oesophageal cancer patients, in which a difference of more than 2% was found for 7/51 weekly evaluations between 3DREM and 4DREM. Three of these evaluations were again for patient 19, in which the target coverage in 4DREM compared to 3DREM was higher in the first week, but lower in week 2 and 3 (Fig. 2B). The largest difference was found in the third week with target coverage of 96.5% in 3DREM and 91.3% in 4DREM, corresponding to scenarios with a MHD ranging from 13.3 to 17.4 Gy. The MHD of the 4DREAL was captured by one SD of both 3DREM and 4DREM for all patients.

## Discussion

In this work, we evaluated the first 20 intra-thoracic indications with large motion treated with IMPT at our proton therapy centre, according to their estimated treatment course dose by our preclinical evaluation method 4DREM, our clinically used 4DREM and the more simple and time-efficient 3DREM. Target coverage was generally sufficient and differences between methods small, especially when accumulated over the entire treatment. The 4DREAL method, based on fraction-specific information, seemed the least conservative assessment method regarding target coverage and homogeneity. Specifically for target coverage, the 4DREAL was the only method where this metric was not extracted from a $V_{w_{\text{min}}}$, as no scenarios including range and setup errors are simulated.

Target coverage was usually better for lung cancer and thymoma patients, relative to oesophageal cancer patients, resulting in a lower number of replannings needed. Anatomical changes resulting in clinical plan adaptation occurred in four oesophageal cases, which corresponded to a $V_{95} < 98\%$ in the weekly evaluation of 3DREM for three cases, in the 4DREM for two cases and the 4DREAL for one case. The weekly 3DREM evaluation possesses the least dose smoothing, as only one CT is considered and no fractionation. Replanning improved target coverage in most cases, especially the following week, but was further fluctuating over the subsequent treatment weeks. However, the inclusion of multiple weekly CTs brought target coverage within clinical limits, as underdosages did not always occur and were not consistently in the same location. Evaluation was always performed on the base-
line CT, also in case of plan adaptation. To investigate the influence of planning on a new CT (new target) and evaluation on the initial planning CT, we additionally evaluated the adapted plans on the corresponding planning CT, but did not find a consistent effect in this set of patients (results can be found in Table Supp.D.

Besides target coverage, the three assessments showed consistent MLD and MHD. The effect of motion did not have a consistent effect on the evaluations. The effect of range and setup errors seemed more impactful, and differences up to 3.1 Gy were observed for the MHD between scenarios. This effect is well understood in case of oesophageal cancer, since the heart is located on the distal end of the target volume and the combination of errors can lead to over- and undershooting.

All three analysed assessments in this study strongly depend on dose warping based on DIR. Large DIR errors can particularly occur when large geometric changes are observed [22]. Ribeiro et al. [23] investigated geometric and dosimetric errors of several DIR methods (including the one used in this study) and found substantial deviations from the ground truth. However, the dosimetric error was reduced when using multiple fields and applying rescanning. Stützer et al. [24] performed different DIR assessments for lung cancer patients and found reliable results, especially within the treatment time frame of five to six weeks.

Rescanned 3D-robustly optimised IMPT plans were clinically suitable for most cases in this patient population with large motion, and could be used to treat them in free breathing. Most centres have restrictions on target motion for proton therapy under free breathing conditions and therefore, limited publications are available on the use of this modality for target volumes with large motion. Motion is underestimated by the mean motion, as different regions of the CTV do not move in the same way or in the same extent. In our clinic, we chose to use the CTV point maximum motion to evaluate the suitability of treating thoracic indications with proton therapy, to overcome potential dose coverage problems in certain areas of the target volume. The motion at the time of the baseline 4DCT was not consistently representative of the motion during treatment, which has also been confirmed by other studies [25]. We generally did not see a correlation of motion amplitude or motion amplitude variation in terms of plan robustness, but certain combinations of motion, anatomical changes, range and setup errors in the scenarios provided worse outcomes. With increased motion, as observed for patient 19, there is a higher chance of failing scenario doses, especially in the weekly 4DREM, which could improve or worsen in combination with a specific range error. 4DREM does not include all possible combinations, which could include or exclude the worst possible combination of errors in terms of target coverage. In the 3DREM approach, target evaluation is based on the ITV, which encompasses the target volume in every single 4DCT phase. Since the target does not reside in the extreme positions at all times, the ITV evaluation could result in an underestimation of the real dose to the moving target.

**Fig. 1.** Target coverage (V95) and target homogeneity index (D2-D98) outcomes for all lung cancer and thymoma patients (patients 1–10) and all oesophageal cancer patients (patients 11–20), according to the three assessment methods. For 3DREM and 4DREM, the V95(Target) is extracted from the resultant Vw_min dose distribution, and the target homogeneity (D2-D98) from all created scenario doses by calculating their mean ± SD.
Daily dose reconstruction and accumulation give us the opportunity to track the accumulated dose along the treatment course. However, current evaluations are based on a weekly 4DCT performed outside the treatment room, which imposes limitations in the current study. The evaluations are performed on limited CT-data captured with the patient immobilized and in treatment position, however this alignment is likely worse than in the treatment room, as no kV imaging or surface alignment are employed. These limitations can be overcome by using daily synthetic 4DCTs (from pre-treatment 4DCBCTs) for dose calculations [26]. Furthermore, we aim to automate the current workflow, as even with computation of only one scenario per fraction in 4DREAL, the process is time and resource consuming.

4DREAL (machine errors, motion and interplay) and 3DREM (setup and range errors) methods are used in our clinic as complementary methods. The 4DREM includes all these effects, and can be used to simulate treatment scenarios to prospectively verify a certain treatment technique before clinical deployment, however is more time intensive and has a limited use of patient-specific data. We can conclude from this study that the combination of 3DREM and 4DREAL provides similar information as 4DREM alone. Furthermore, the effects of anatomical changes, range and setup errors have more impact than the effects of motion, as the 3DREM target coverage results were consistently worse than the 4DREAL results. This highlights the importance to include periodical verification imaging in evaluation of robustness, in which at least weekly imaging is recommended to not underestimate the dose smoothening effect. The results obtained for 3DREM were similar to 4DREM, demonstrating the potential to make plan evaluation simpler and more time-efficient. Nevertheless, we need to be careful to conclude that an averaged 4DCT-based evaluation without the consideration of interplay effects is sufficient for thoracic indications. That the 3DREM and 4DREM gave similar outcomes in this study could be dependent on how the treatment plans were generated (robust optimisation, spot size) and delivered (dose rate, repaint). Therefore, we encourage careful evaluation with any deviations from our protocol, in order to be able to accurately and comprehensively assess all effects that could disturb adequate
treatment delivery. Further work is warranted to merge all aspects of the 4D evaluation in one comprehensive and efficient method within the clinical setting.

Conclusion

By assessment of the treatment course dose using 3DREM, 4DREM and 4DREAL, we found adequate robustness in thoracic patients with large motion treated with conventional fractionated IMPT in free breathing. With weekly plan evaluation and offline adaptation, adequate target coverage and stable OAR dose was demonstrated at course-wise dose level independent of the method used for dose assessment. Machine errors and motion with interplay effects seemed not to considerably impact target dose coverage. On the other hand, anatomical variations, setup and range errors had a more substantial influence on IMPT dose deposition for intra-thoracic tumours with large motion. This emphasizes the necessity of repeated verification imaging, to be able to detect anatomical changes and their impact on the dose distribution.

Conflict of Interest

Dr. Langendijk reports personal fees from IBA, other from IBA, other from Philips, other from MIRADA, other from RaySearch, other from Siemens, other from Elekta, other from Leonie, outside the submitted work. As of 01/09/2021 C. Oraboni Ribeiro is full-time employee of Ion Beam Applications S.A. (IBA). This study has been performed prior to that. No financial or in-kind contributions from IBA have been received. All other authors have no conflicts of interest to disclose.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109575.

References


