Physics in Medicine & Biology

IPEM
Institute of Physics and
Engineering in Medicine

PAPER

Dosimetric comparison of autocontouring techniques for online adaptive proton therapy

A Smolders[1,2] , E Choulilitsa[1,2], K Czerska[1], N Bizzocchi[1], R Krcek[1,4], A Lomax[1,2], D C Weber[1,3,4] and F Albertini[1]

1   Paul Scherrer Institute, Center for Proton Therapy, Switzerland
2   Department of Physics, ETH Zurich, Switzerland
3   Department of Radiation Oncology, University Hospital Zurich, Switzerland
4   Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Switzerland

E-mail: andreas.smolders@psi.ch

Abstract

Objective. Anatomical and daily set-up uncertainties impede high precision delivery of proton therapy. With online adaptation, the daily plan is reoptimized on an image taken shortly before the treatment, reducing these uncertainties and, hence, allowing a more accurate delivery. This reoptimization requires target and organs-at-risk (OAR) contours on the daily image, which need to be delineated automatically since manual contouring is too slow. Whereas multiple methods for autocontouring exist, none of them are fully accurate, which affects the daily dose. This work aims to quantify the magnitude of this dosimetric effect for four contouring techniques. *Approach.* Plans reoptimized on automatic contours are compared with plans reoptimized on manual contours. The methods include rigid and deformable registration (DIR), deep-learning based segmentation and patient-specific segmentation. *Main results.* It was found that independently of the contouring method, the dosimetric influence of using *automatic* OAR *contours* is small (<5% prescribed dose in most cases), with DIR yielding the best results. Contrarily, the dosimetric effect of using the *automatic target contour* was larger (>5% prescribed dose in most cases), indicating that manual verification of that contour remains necessary. However, when compared to non-adaptive therapy, the dose differences caused by automatically contouring the target were small and target coverage was improved, especially for DIR. *Significance.* The results show that manual adjustment of OARs is rarely necessary and that several autocontouring techniques are directly usable. Contrarily, manual adjustment of the target is important. This allows prioritizing tasks during time-critical online adaptive proton therapy and therefore supports its further clinical implementation.

1. Introduction

Proton therapy results in a lower integral dose and improved organ-at-risk (OAR) sparing compared to photon therapy for the same target dose because of the peaked depth-dose profile of particles (Paganetti 2012). This advantage is however conditional on accurate positioning of the dose peak, which is sensitive to the tissue densities along the beam path (Lomax 2008, Zhang *et al* 2011). The dose peak position is therefore affected by daily set-up variations and anatomical changes, such as weight loss and tumor shrinkage, which impede high precision delivery of proton therapy. Conventionally, anatomical and set-up uncertainty is managed by increasing treatment plan robustness, either by robust optimization (Liu *et al* 2012, Unkelbach *et al* 2018) or adding margins around the clinical target volume (CTV) (Albertini *et al* 2011). In either case, the dose to the healthy tissue increases and the advantage of proton therapy decreases.

Instead of increasing the robustness, online adaptive proton therapy aims to reduce the aforementioned uncertainty (Albertini *et al* 2020, Paganetti *et al* 2021). This alleviates the need for robustness and retains the

advantage of proton therapy. By acquiring a 3D image in treatment position shortly before the delivery and adapting the treatment plan using this image, both anatomical and set-up uncertainty are substantially reduced. However, if the time between imaging and delivery is long, these uncertainties increase again because of potential patient movement and slow intrafractional anatomical changes (e.g. bladder filling, organ drift). In addition, increased technical and human resources have to be taken allocated due to longer time slots for treatment. Therefore, all adaptation processes need to be as fast as possible, which additionally improves patient comfort because of reduced overall treatment time.

Online adaptation requires OAR and target contouring on the new images, plan reoptimization and quality assurance (QA). The most time-consuming adaptation process is contouring (Lim-Reinders *et al* 2017), which can be sped up by automation in several ways, such as rigid registration (RR), deformable registration (DIR), deep learning based autosegmentation or patient-specific segmentation (PSS) (Smolders *et al* 2023). A recent comparison found that DIR and PSS reached the highest accuracy for a large set of OARs and target volumes for patients with head-and-neck and lung cancer (Smolders *et al* 2023). The evaluation was however limited to geometrical measures such as dice score and Hausdorff distance (Taha and Hanbury 2015). Despite the promising results, the automatic contours were not perfectly corresponding to manually drawn ones, and the dosimetric impact of these inaccuracies was not assessed. Furthermore, previous work has shown that geometrical measures do not correlate well with dosimetric differences (Tsuji *et al* 2010, Voet *et al* 2011, Kaderka *et al* 2019, Sherer *et al* 2021). As a consequence, it is unclear whether the proposed methods can directly be used in adaptive therapy.

In this work, the dosimetric influence of using automatically generated contours in online adaptive proton therapy was assessed. The analysis included patients with head and neck cancer (HNC) and non-small cell lung cancer (NSCLC), therefore covering a wide range of anatomical regions and deformations relevant for adaptive proton therapy. Four autocontouring methods, including RR, deformable registration, deep learning based autocontouring and patient-specific neural networks were compared and the necessity of manual adjustments was evaluated. The rest of this paper is organized as follows: section 2 introduces the methods for contour propagation, the datasets and the metrics used for evaluation. The results are stated in section 3, followed by a discussion in section 4 and conclusion in section 5.

2. Materials and methodology

2.1. Automatic contouring methods
The comparison includes four methods to automatically contour the daily CT. They are only shortly summarized hereafter, as these methods were described in detail previously (Smolders *et al* 2023).

(i) Rigid registration (RR): in RR, the planning CT is translated and rotated to match the position of the daily CT. This method is fast, consistent and yields good results if the anatomy of the patient is not deforming substantially. The same transformation is then applied to the planning contours to obtain the daily contours. RR was implemented with `elastix` (Klein *et al* 2010).

(ii) Deformable registration (DIR): in case of anatomical deformations, DIR can be used to obtain a deformable mapping between the planning CT and the daily CT, which is subsequently used to deform the planning contours into the daily ones. DIR is slow compared to RR, and different implementations usually lead to inconsistent results (Brock *et al* 2017). Moreover, DIR algorithms generally fail to model anatomical transformations that involve non-smooth deformations, such as the formation or removal of mass or the sliding of tissue boundaries. Here, a b-spline DIR from `plastimatch` was used (Sharp *et al* 2010).

(iii) Commercial deep-learning segmentation: deep convolutional neural networks (CNNs) are trained on a large dataset with example images and contours. They can be used directly to predict the contours of a new image. CNNs are generally fast and accurate, but may fail if the anatomy of the patient under study is significantly different from the training images. We used the software Limbus Contour 1.7 (AI Limbus Inc., 2076 Athol Street, Regina, SK S4T 3E5, Canada).

(iv) Patient-specific segmentation (PSS): patient specific CNNs are a type of CNNs which are trained specifically to segment the contours of one patient. In adaptive therapy, they are trained on the planning CT. Similar to normal CNNs, they are fast, with the additional advantage that they have been trained on the planning CT which is highly similar to the subsequent daily CTs. The implementation details can be found in Smolders *et al* (2023).

Figure 1. The 3-field NSCLC (left) and HNC (right) reference plans on the planning CTs.
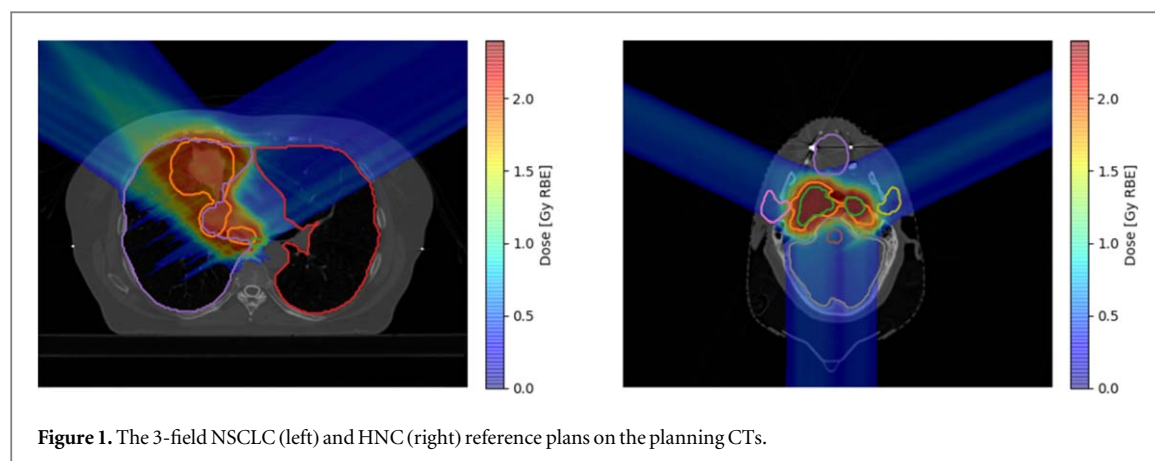
2.2. Patient data

The comparison includes 10 patients: 5 with NSCLC and 5 with various types of HNC. The 5 NSCLC *patients* have previously been described for other purposes in Josipovic *et al* (2016), Nenoff *et al* (2020), Amstutz *et al* (2021), Nenoff *et al* (2021). For each patient, 9 repeated CTs were acquired on 3 separated days with the same CT scanner as the planning CT, each day consisting of 3 repeated CTs. Here, each repeated scan is considered to represent a single fraction of an online adaptive treatment. The *HNC patients* were treated with proton therapy at Center for Proton therapy (CPT) and, as per clinical protocol, underwent 4–7 control CTs during the course of the treatment with the same CT scanner as the planning CT. These control CTs can be considered as samples of daily CTs from an online adaptive treatment, as they contain both daily and longitudinal anatomic and set-up changes. The repeated CTs of all 10 patients were retrospectively recontoured slice-wise by expert radiation oncologists according to clinical protocols with the aid of Eclipse (Varian Medical Systems, Palo Alto, CA, USA) (Nenoff *et al* 2021).

2.3. Reference treatment plans

For each patient, a clinically acceptable proton radiotherapy plan was retrospectively designed and optimized on the planning CT using the in-house treatment planning software FIonA. For the NSCLC patients, the plans consist of 3 fields (figure 1) delivering a homogeneous 2 Gy RBE fraction dose to the PTV (Nenoff *et al* 2021), which was generated by 2 mm isotropic expansion of the CTV. This margin is close to values for adaptive proton therapy reported in recent literature (Jagt *et al* 2018, Nenoff *et al* 2019, Berger *et al* 2021, Nenoff *et al* 2021), and significantly smaller than the current clinical standard because the online adaptation reduces anatomic and set-up uncertainty (van Kranen *et al* 2016, Sonke *et al* 2019). Additionally, $\pm 3\%$ range uncertainty was considered in the robust optimization. No other robustness was included in the optimization. The field directions were chosen for each patient individually to maximize OAR sparing and the spot-weights were optimized using a multi-field optimization (Matter *et al* 2019), following clinical practice. Each plan was reviewed by a radiation oncologist. Dose constraints were defined for the heart, esophagus, spinal cord and lungs. In order to highlight the effect of different OAR contours on the dose distribution, it was ensured that each constraint was affecting the plan, i.e. the clinical constraint was sometimes tightened for the purpose of this study so that its enforcement affected the dose distribution.

The HNC plans deliver 1.6 Gy RBE per fraction to the low-dose PTV with a simultaneous integrated boost of 2.2 Gy RBE to the high-dose PTV (Wu *et al* 2003). All plans use the same 3-field configuration with gantry angles $65°$, $180°$ and $295°$ and were optimized with (figure 1). The constraints were chosen in line with the standard clinical goals for HNC and were, contrary to the NSCLC cases, the same for all patients. Also here, the PTV was defined by a 2 mm isotropic expansion of the CTV and $\pm 3\%$ range uncertainty was considered.

2.4. Compared adaptation scenarios

Following the adaptive workflow described in Matter *et al* (2020), the reference plan was reoptimized on each daily CT to simulate daily adaptation. The number of fields and their beam angles were kept constant, and spot positions and weights were fully reoptimized. Additionally, optimization constraints and their respective weights remained constant, as well as the optimization grid resolution and location. This reoptimization was repeated in several settings for comparison, each of which is detailed below.

- Manually contoured plans: these plans are reoptimized using all the manual contours. This represents the ideal situation of adaptation where the contours adhere to the current clinical standard. Although this is not

feasible in routine clinical care due to time constraints, these plans represent the ideal scenario to which all other plans can be compared.

- Automatically contoured plans: these plans are reoptimized using the propagated contours for each autocontouring method. We refer to these plans with respect to the propagation method, e.g. DIR *contoured plan*, representing a plan optimized on contours propagated through the DIR algorithm. To separate the effects of OAR segmentation from target volume segmentation, these *automatically contoured plans* are reoptimized in two settings:

    (i) Optimized with the automatic OARs and manual target volume: in this case, the plan is optimized with the *automatic OAR contours* and *manually delineated target volume*, so that the effect of using automatic OAR contours is isolated. This approach represents an online workflow in which OARS are automatically contoured without clinician intervention, whereas the target volume is manually adjusted. This step already reduces the contouring time significantly compared to adjusting all propagated contours.

    (ii) Optimized with the automatic OARs and automatic target volume: in this case, the plan is optimized solely on *automatic contours*, which leads to the largest time gain. Note that target volume segmentation is not included in the commercial segmentation tool, so this technique is omitted for this case.

- Non-adaptive plans: these plans are the reference treatment plans, optimized for each patient on the planning CT, but recalculated on the daily anatomy. This is the current clinical standard without adaptation, and it offers a baseline method for comparison. The daily anatomy is rigidly registered to the planning CT before plan recalculation, which mimics 3D image-based patient positioning through a couch shift and rotation, using 6 degrees of freedom.

- Logfile back calculated plans: all plans based on manually delineated contours were delivered to a phantom on Gantry 2 at CPT (Pedroni *et al* 2004). The dose was then reconstructed using the machine log files, including inaccuracies in spot, couch and gantry positioning (Scandurra *et al* 2016). The dose difference between these plans and the manual plans allows quantification of the delivery accuracy, which can be compared to the other dose differences. This back calculation does not account for range uncertainty or patient shifts between the CT acquisition and the treatment, therefore only representing an upper bound of the delivery accuracy.

Even though the comparison only includes 10 patients, each patient has multiple repeated CTs, yielding 45 plans for NSCLC and 28 for HNC per method. These 73 plans were repeated for the 4 automatic contouring methods, once with manual and once with automatic target, requiring a total of 657 plan optimizations. Additionally, all 73 plans based on manual contours were delivered once for log-file back calculation.

2.5. Evaluation metrics

To evaluate the dosimetric influence of using automatic contours, each *automatically contoured plan* was compared to the respective *manually contoured* one. The *automatically contoured plan* yielding the smallest dose difference with the *manually contoured plan* is assumed to be the best, as it approximates closest the clinically ideal situation. The dose difference was evaluated by calculating the voxel-wise absolute dose difference between the two plans, and creating dose-difference-volume-histograms (DDVH), analogous to the calculation of a DVH but with the absolute dose difference instead of the dose. This allowed to calculate the dose difference metric DDx for a structure. The DDx indicates that the dose difference in $(100 - x)$% of the volume is lower than its value, e.g a DD5 of 7% means that 95% of the volume receives a dose difference smaller than 7% of the prescribed dose. In case the plan is reoptimized on the manual target, x was set to 2%. For the propagated target $x = 5$% because of the larger dosimetric differences. The DDVHs were created using the manual structure volumes, as these are considered to be the ground-truth structures.

Paired Wilcoxon signed-rank tests were performed between all methods to test which one leads to significantly lower dose differences. Since the DDx of different organs cannot be considered independent, the average DDx was first calculated over all organs, and the test was performed on this average.

Whereas the DDx allows to compare the dosimetric influence of using automatic contours, it does not directly link to a clinically meaningful metric. Therefore, we further calculated for each *automatically contoured plan* and each OAR either D_{max} or D_{mean}, depending on the clinical relevance, and compared it to the respective value in the *manually contoured plan*. For the targets, D98 and V95 were evaluated. The differences with the *manually contoured plan* indicate how much this clinically relevant parameter might vary due to automatic contouring for a single fraction. As for the DDx, the DVHs were calculated using the manual structures, i.e. the differences in D_{max} or D_{mean} are solely due to the difference in dose distribution.
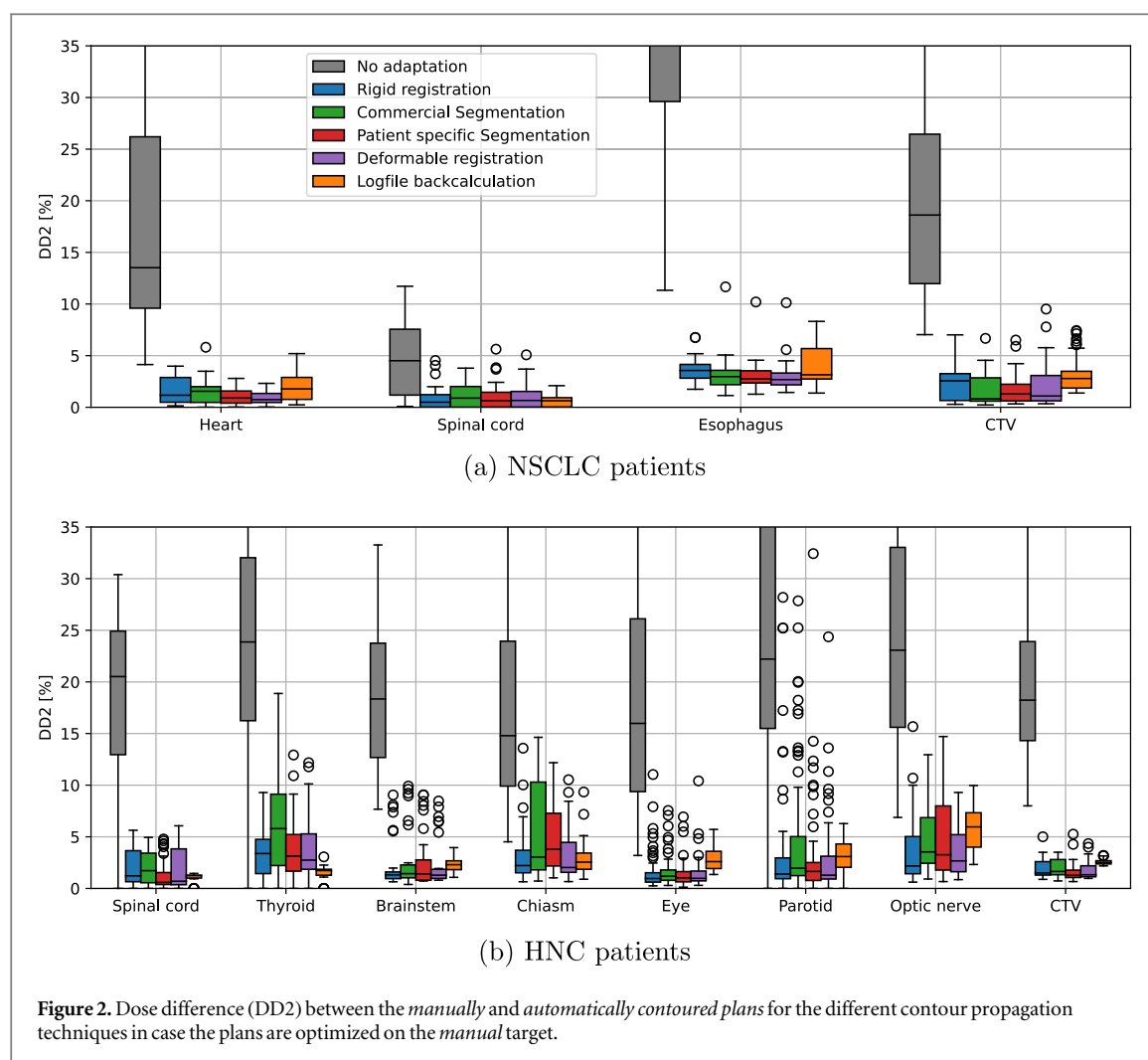
Figure 2. Dose difference (DD2) between the *manually* and *automatically contoured plans* for the different contour propagation techniques in case the plans are optimized on the *manual* target.

3. Results

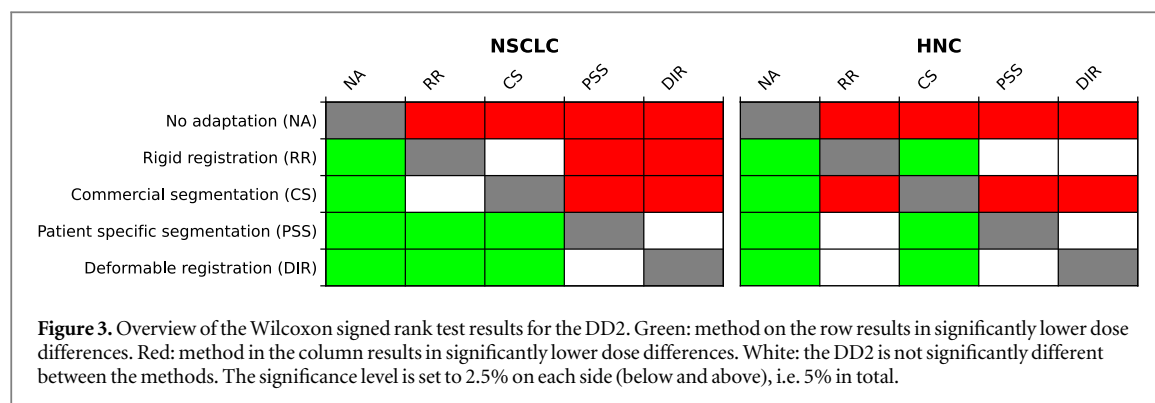## 3.1. Plans optimized on automatic OARS and manual target volume

Following an adaptive workflow where only the target is manually adjusted by a clinician, the dose differences with the *manually contoured plan* are small compared to those for conventional therapy without adaptation (figure 2). Indeed, independently of the contouring method, adapting the treatment leads to significantly lower dose differences than not adapting (figure 3). Also the respective clinically relevant parameters are affected less when adapting with automatic contours than without adaptation (figure A1).

The dosimetric differences with the *manually contoured plans* are also small in absolute value for most methods and organs: the DD2 for 87% of all cases below 5%. This means that 98% of the volume of those organs receive a dose difference less than 5% of the prescribed dose, and this difference is usually much smaller. For DIR alone, this holds in 92% of the cases. The dose differences are in 2/3 of the cases even smaller than the difference with the back calculated plan, i.e. smaller than the delivery accuracy. Similarly, using automatic OAR contours, D_{mean} and D_{max} remain in 94% of all cases within $\pm 5\%$ of the respective value when using all manual contours (figure A1).

The dose difference is also small for some automatic OAR contours that were not geometrically accurate. For example, the median esophagus dice was only 0.69 for RR (Smolders *et al* 2023), but the corresponding median DD2 is 3.6%. This implies that approximate contours are often sufficient for reoptimization.

For the parotid glands, the DD2 is sometimes larger than 20%, much larger than the delivery accuracy. Also the difference in D_{mean} is sometimes much larger than for the other OARs. This is due to several reasons. Firstly, the parotids are located close to the target volume and affect the dose distribution significantly, so that any geometric deviation influences the reoptimized plan. Secondly, they are in general difficult to segment on CT because of the poor soft tissue contrast. Lastly, the parotids often move and change volume and shape throughout the treatment (Ricchetti *et al* 2011), making them more difficult to automatically segment. To a lesser extent, this is also true for the thyroid and chiasm.

Figure 3. Overview of the Wilcoxon signed rank test results for the DD2. Green: method on the row results in significantly lower dose differences. Red: method in the column results in significantly lower dose differences. White: the DD2 is not significantly different between the methods. The significance level is set to 2.5% on each side (below and above), i.e. 5% in total.

Because the dose differences are mostly small independently of the contouring method, all could be used in online adaptive therapy to clinical advantage. However, some methods lead to better results than others (figure 3). In the head region, RR performs better than the commercial segmentation, but there is no significant difference in the thorax region, although generally DIR and PSS perform best. In the head and neck region, RR is a valid alternative for both.

3.2. Plans optimized on automatic OARS and automatic target

In case the treatment is reoptimized based on the propagated target rather than manually delineated targets, much larger dosimetric differences are found (figures 4 and A2). In more than 70% of all cases the DD5 is above 5%. The dosimetric differences in the OARs can be as large as for conventional proton therapy without adaptation, and are much larger than the difference with the back calculated plans. It is important to note that all these differences are due to the propagation of the target and not the OARs, as shown before (figure 2). For the target itself, the differences are still much smaller than with non-adapted therapy, highlighting the advantage of adaptation even when the target is automatically delineated. For one patient in the NSCLC data, the PSS performs very badly, due to significant anatomical deformation of the tumor (see Smolders *et al* (2023)). The Wilcoxon rank test found however that adapting still results in significantly lower DD2 than not adapting (results not shown). However, it did not find significant differences between the methods, except that the PSS performed worse than DIR for the NSCLC data.
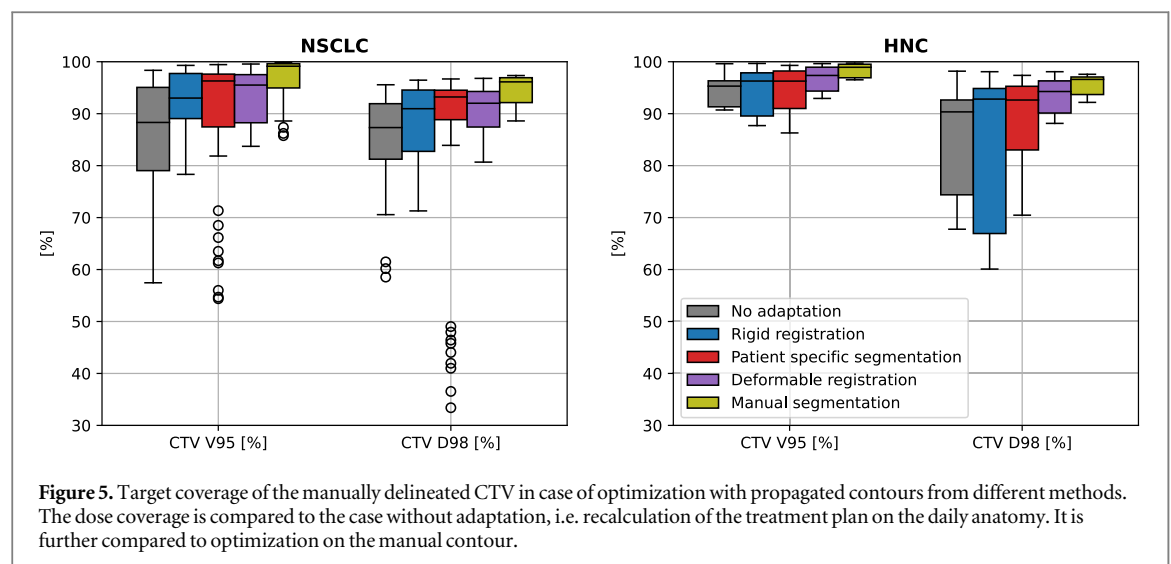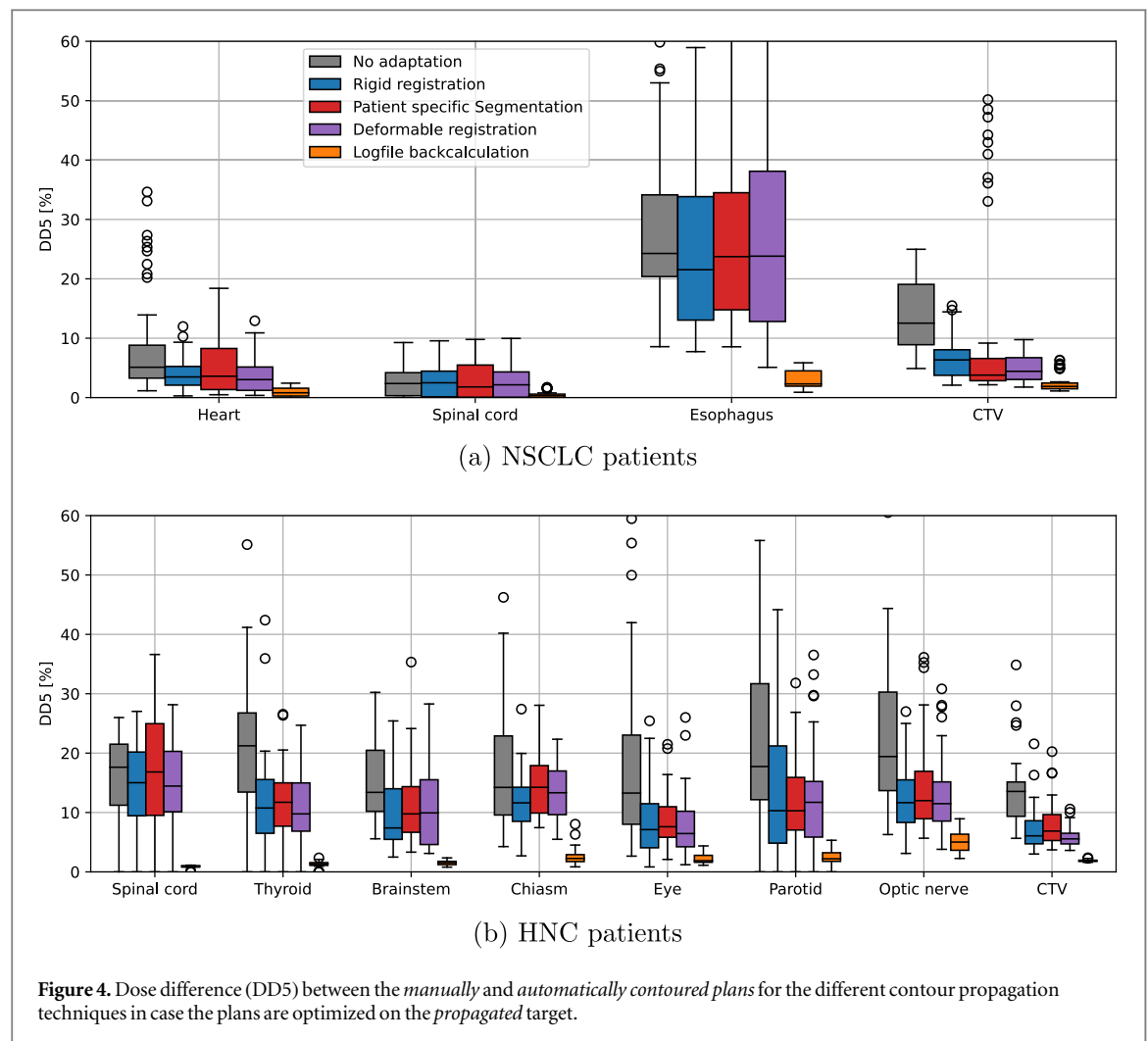
Because propagating the target clearly influences the dose distribution, it is important to evaluate target coverage for each propagation technique (figure 5). The target coverage without adaptation is in some cases very low (V95% < 80% prescribed dose) and adaptation generally leads to higher coverage, even though the difference is not always significant (figure 6). RR and PSS lead to similar coverage, and are both significantly outperformed by DIR.

Whereas previous results indicate that RR is a valid alternative for contouring both OARs and targets in the head and neck region, it leads to large degradation in target coverage for some cases (figure 5). Similarly, for one patient with NSCLC, PSS performs very poorly. Therefore, routine manual verification would be necessary to ensure that clinical goals are met.
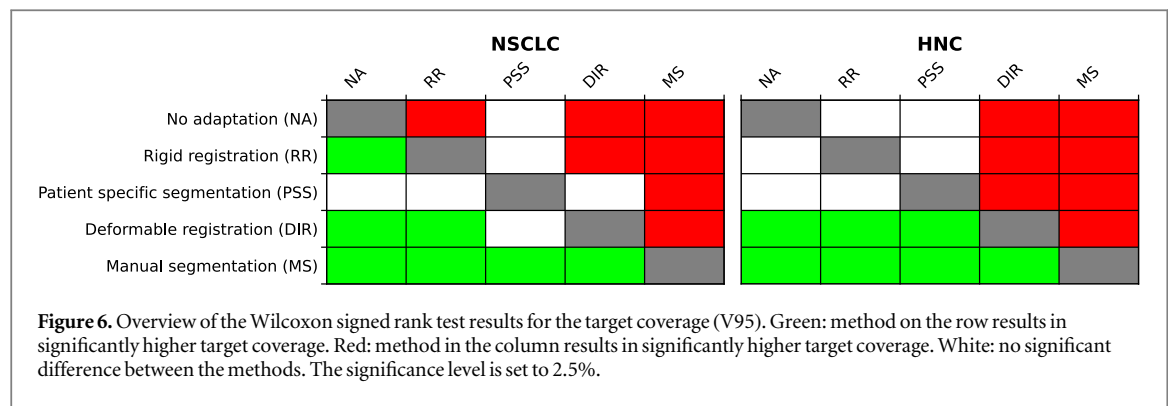
4. Discussion

Our results clearly highlight the advantage of adapting and reoptimizing the treatment plan on the daily anatomy for proton therapy. Indeed, even without manual intervention and correction of daily contours, the benefit of plan reoptimization is evident, both regarding target coverage and dosimetric differences in the OARs. Already using simple contour propagation methods, such as RR, improves the results compared to non-adaptive treatments in nearly all cases. More advanced techniques like deformable registration further enhance the benefit of adaptation.

The analysis shows that reoptimization on automatic OAR contours generally leads to very small dose differences with reoptimization on manual contours, as in IMRT (Feng *et al* 2012, Guo *et al* 2021, Zhang *et al* 2022) and VMAT (van Rooij *et al* 2019), even if these contours are not very accurate. Nevertheless, in some cases, the dose difference is still large. This indicates that manual adjustments are not necessary for most OARs during adaptive proton therapy of HNC and NSCLC and that only a subset of structures requires inspection and manual adjustment before starting a daily optimization, similar to findings for prostate cancer (Cao *et al* 2020). Regarding the different OARs, dose differences larger than 10% DD2 were found in the parotid glands and thyroid, even for the best contouring method. Therefore, considering the time limits for manual intervention, review of these contours should be prioritized. Additionally, QA checks of the daily contours can be employed to

(a) NSCLC patients

(b) HNC patients

Figure 4. Dose difference (DD5) between the *manually* and *automatically contoured plans* for the different contour propagation techniques in case the plans are optimized on the *propagated* target.

Figure 5. Target coverage of the manually delineated CTV in case of optimization with propagated contours from different methods. The dose coverage is compared to the case without adaptation, i.e. recalculation of the treatment plan on the daily anatomy. It is further compared to optimization on the manual contour.

prioritize contour inspection. This work shows that different independent methods can be used to obtain daily contours, and comparing the contours of different methods might allow to identify inaccuracies. In a follow-up work, the results of this study will be used to assess whether and how such a QA check could be used in an online adaptive workflow.

Even if time limitations do not permit manual inspection of all contours before the fraction is delivered, inspection after delivery remains a viable option in normofractionated irradiation. Such an offline review is less

	NSCLC NA	NSCLC RR	NSCLC PSS	NSCLC DIR	NSCLC MS	HNC NA	HNC RR	HNC PSS	HNC DIR	HNC MS
No adaptation (NA)	gray	red	white	red	red	gray	white	white	red	red
Rigid registration (RR)	green	gray	white	red	red	white	gray	white	red	red
Patient specific segmentation (PSS)	white	white	gray	white	red	white	white	gray	red	red
Deformable registration (DIR)	green	green	white	gray	red	green	green	green	gray	red
Manual segmentation (MS)	green	green	green	green	gray	green	green	green	green	gray

Figure 6. Overview of the Wilcoxon signed rank test results for the target coverage (V95). Green: method on the row results in significantly higher target coverage. Red: method in the column results in significantly higher target coverage. White: no significant difference between the methods. The significance level is set to 2.5%.

constrained by time limitations and a more elaborate inspection is possible. Even though it cannot alter the delivered dose, it can trigger adaptation in the subsequent fractions in case an inaccuracy in contours caused significant dose difference.

Contrary to using the automatic OAR contours, reoptimizing the plan on the propagated target instead of the manual one clearly influences the dose. Similar results were found for IMRT (Tsuji *et al* 2010). Not only is the target coverage adversely affected, but also the dose to the OARs is modified substantially. This indicates that manual verification and adjustment of the target contours is important and should be included in the online adaptive workflow. To speed up such adjustments, the target contours can be modified first in offline review. These offline adapted contours can be used as the reference during propagation, reducing the amount of online manual adjustments.

The difference in D_{max} or D_{mean} allows to interpret the clinical relevance of a dosimetric change, but this metric should be interpreted with care. Firstly, these are fractional differences, so a positive difference in one fraction can be canceled out by a negative difference in another one. Secondly, a reduction in D_{max} or D_{mean} when using autocontouring should not be interpreted as an improvement of the plan. Such reduction can e.g. be caused by an overly large OAR contour, which would indeed lead to a reduction of the dose inside the OAR, but would also cause a reduction in target coverage.

The results found in this study are specific to the treatment location (head and neck; lung), geometry (i.e. field angles) and optimization constraints. For example, we found that the dose is sensitive to the shape of the parotids, partly because they are close to the tumor and because the constraint is heavily affecting the dose. This might not be true for different indications, treatment geometries or optimization constraints. Therefore, this analysis should be repeated when prioritizing OAR inspection for adaptive therapy of other indications. Alternatively, it could be performed for each patient separately, either by perturbing and deforming the contours on the planning CT and evaluating the dosimetric influence, or during offline review after a few fractions to speed up the process in the remaining days.

Of all contouring methods, DIR generally exhibits the most promising results. However, as stated before, DIR can fail in case of formation or removal of mass or sliding tissue boundaries. The lung CTs here were acquired in deep inspiration breath hold, and therefore suffer only slightly from the sliding boundary issue. Further, in this limited patient cohort, only a few patients exhibited strong tumor shrinkage. Therefore, future work should study specifically cases where DIR might fail, as other methods might be preferred there.

Additionally, running the DIR algorithm for these cases takes on average 2.5 min, which is more than 5 times longer than any of the other methods. In view of time, it could be beneficial to run one of the other methods instead. As the influence of using the propagated OAR contours is anyway small, using another method than DIR is likely sufficient. Furthermore, the target contour will need to be checked for any method. Even though the target contour from RR or PSS would likely require more manual corrections, the combined time difference could still be positive so that using another method than DIR would lead to a time benefit.

The repeated images in this study are acquired with the same CT scanner as the planning CT, mimicking an online adaptive workflow with in-room CT. However, CBCT or MR based daily adaptation is gaining interest, mainly because of the presence of gantry-mounted CBCTs and superior soft tissue contrast in MRI (Albertini *et al* 2020). Even though not all contouring techniques described here are directly suitable for CBCT and MR, the conversion of these images into pseudo-CTs remains necessary for dose recalculation, and these pseudo-CTs can be used for contouring.

An important limitation of the study is that all automatic plans are compared to the treatment plan optimized on the manual contours, but that these manual contours themselves are also subject to variations. Indeed, many other studies have shown that the inter-observer contour variability can be substantial for both

NSCLC and HNC (Deeley *et al* 2011, Brouwer *et al* 2012, Mattiucci *et al* 2013, Verhaart *et al* 2014, Tao *et al* 2015, Yang *et al* 2018, van der Veen *et al* 2019, Wong *et al* 2020). Therefore, in the case that the dosimetric differences are small, the plan optimized on the propagated contours can be just as valid as that calculated to the manual contours. Contrarily if the differences are large, the manual plan can be assumed better because the contours were verified by expert clinical personnel.

5. Conclusion

In this work, different methods for automatic contouring in online adaptive proton therapy were compared dosimetrically. We found that the influence of reoptimizing daily plans on *automatic OAR contours* instead of *manual contours* is small, independently of the contouring method. This means that multiple techniques are usable and that manual adjustments are only rarely necessary. Contrarily, propagating the target with any method can significantly alter the dose in the OARs and adversely affect target coverage, therefore pointing out the importance of manual verification of the target contours. Overall, deformable registration yielded the highest target coverage and lowest dose differences in the OARs.
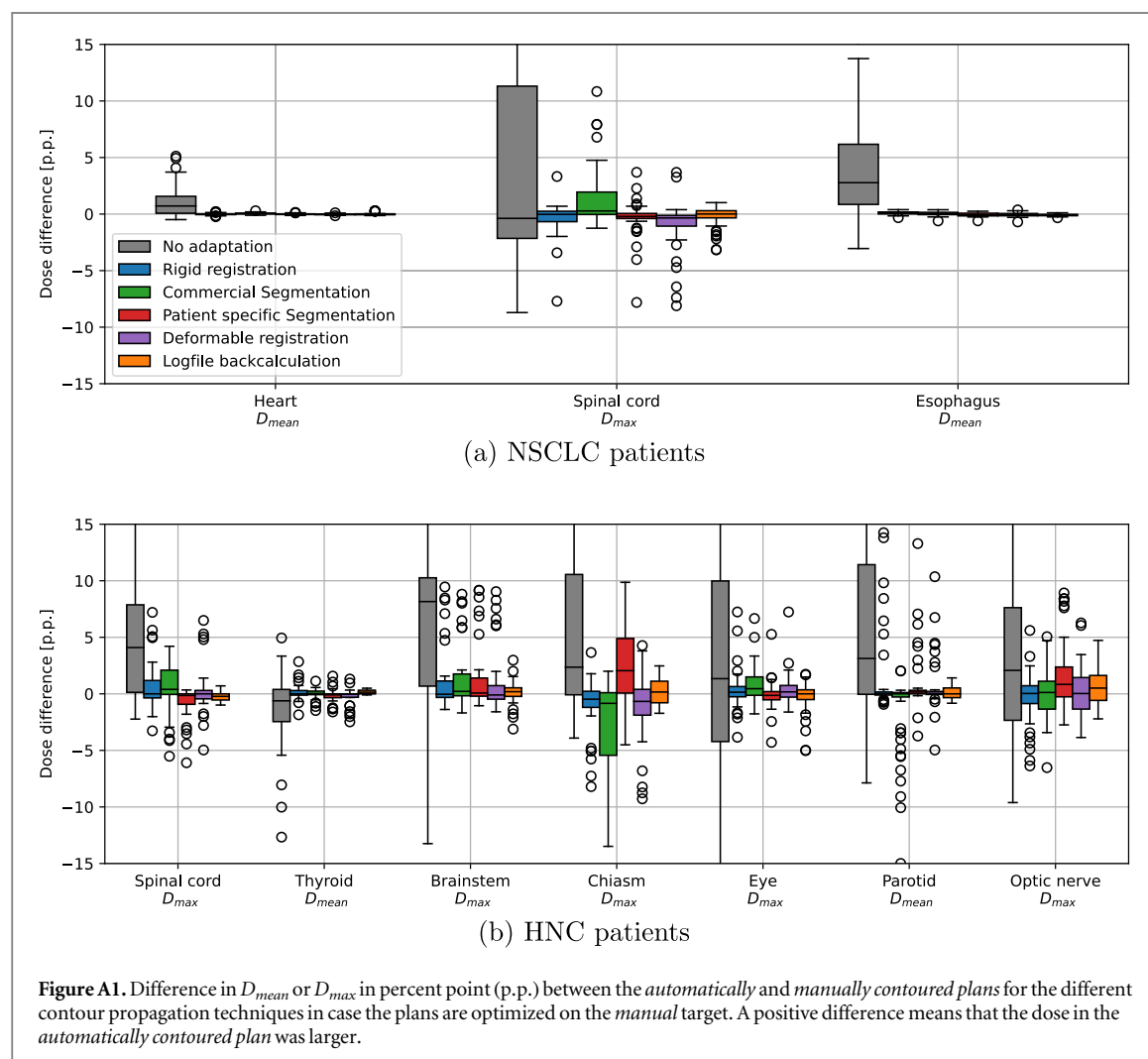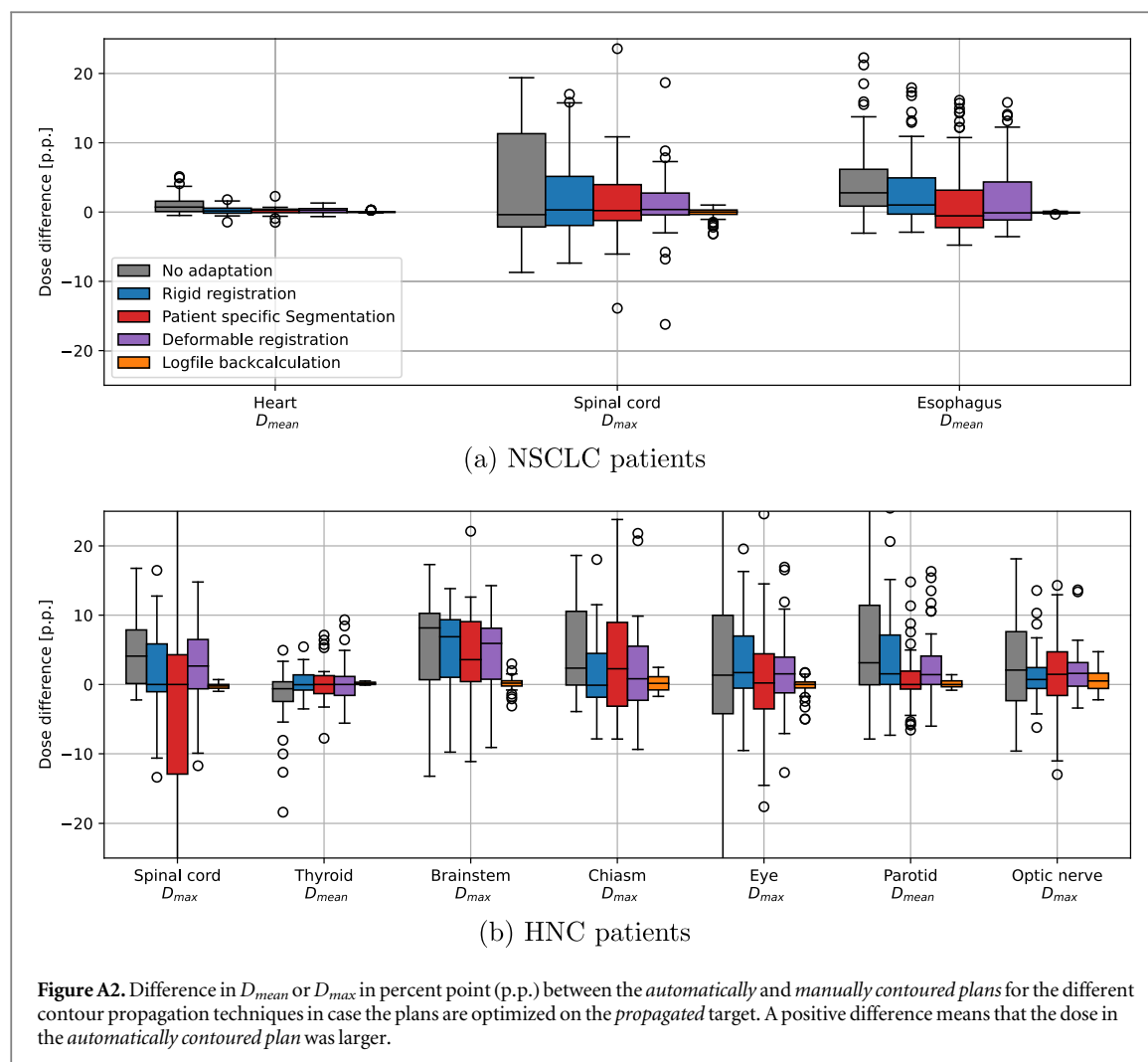
Acknowledgments

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Appendix. Differences in clinically meaningful metrics

(a) NSCLC patients

(b) HNC patients

Figure A1. Difference in D_{mean} or D_{max} in percent point (p.p.) between the *automatically* and *manually contoured plans* for the different contour propagation techniques in case the plans are optimized on the *manual* target. A positive difference means that the dose in the *automatically contoured plan* was larger.

(a) NSCLC patients

(b) HNC patients

Figure A2. Difference in D_{mean} or D_{max} in percent point (p.p.) between the *automatically* and *manually contoured plans* for the different contour propagation techniques in case the plans are optimized on the *propagated* target. A positive difference means that the dose in the *automatically contoured plan* was larger.

ORCID iDs

A Smolders ⬤ https://orcid.org/0000-0003-2874-3634

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

Albertini F, Matter M, Nenoff L, Zhang Y and Lomax A 2020 Online daily adaptive proton therapy *Br. J. Radiol.* **93** 20190594

Amstutz F, Nenoff L, Albertini F, Ribeiro C O, Knopf A C, Unkelbach J, Weber D C, Lomax A J and Zhang Y 2021 An approach for estimating dosimetric uncertainties in deformable dose accumulation in pencil beam scanning proton therapy for lung cancer *Phys. Med. Biol.* **66** 105007

Berger T *et al* 2021 Dosimetric impact of intrafraction motion in online-adaptive intensity modulated proton therapy for cervical cancer *Int. J. Radiat. Oncol.***Biol.***Phys.* **109** 1580–7

Brock K K, Mutic S, McNutt T R, Li H and Kessler M L 2017 Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy committee task group no. 132: report *Med. Phys.* **44** e43–76

Brouwer C L *et al* 2012 3D variation in delineation of head and neck organs at risk *Radiat. Oncol.* **7** 32

Cao M, Stiehl B, Yu V Y, Sheng K, Kishan A U, Chin R K, Yang Y and Ruan D 2020 Analysis of geometric performance and dosimetric impact of using automatic contour segmentation for radiotherapy planning *Frontiers Oncol.* **10** 1762

Deeley M A *et al* 2011 Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study *Phys. Med. Biol.* **56** 4557–77

Feng M, Demiroz C, Vineberg K A, Eisbruch A and Balter J M 2012 Normal tissue anatomy for oropharyngeal cancer: contouring variability and its impact on optimization *Int. J. Radiat. Oncol.***Biol.***Phys.* **84** e245–9

Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z and Hu W 2021 The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer *Radiat. Oncol.* **16** 113

Jagt T, Breedveld S, van Haveren R, Heijmen B and Hoogeman M 2018 An automated planning strategy for near real-time adaptive proton therapy in prostate cancer *Phys. Med. Biol.* **63** 135017

Josipovic M, Persson G F, Dueck J, Bangsgaard J P, Westman G, Specht L and Aznar M C 2016 Geometric uncertainties in voluntary deep inspiration breath hold radiotherapy for locally advanced lung cancer *Radiother. Oncol.* **118** 510–4

Kaderka R *et al* 2019 Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients *Radiother. Oncol.* **131** 215–20

Klein S, Staring M, Murphy K, Viergever M A and Pluim J P W 2010 elastix: a toolbox for intensity-based medical image registration *IEEE Trans. Med. Imaging* **29** 196–205

Lim-Reinders S, Keller B M, Al-Ward S, Sahgal A and Kim A 2017 Online adaptive radiation therapy *Int. J. Radiat. Oncol.***Biol.***Phys.* **99** 994–1003

Liu W, Zhang X, Li Y and Mohan R 2012 Robust optimization of intensity modulated proton therapy *Med. Phys.* **39** 1079–91

Lomax A J 2008 Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions *Phys. Med. Biol.* **53** 1043–56

Matter M, Lomax A J, Albertini F, Sonke J-J and Vaterlaus A 2020 Technical implementation of daily adaptive proton therapy *PhD thesis* ETH Zurich

Matter M, Nenoff L, Meier G, Weber D C, Lomax A J and Albertini F 2019 Intensity modulated proton therapy plan generation in under ten seconds *Acta Oncol.* **58** 1435–9

Mattiucci G C *et al* 2013 Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? *Acta Oncol.* **52** 1417–22

Nenoff L *et al* 2020 Deformable image registration uncertainty for inter-fractional dose accumulation of lung cancer proton therapy *Radiother. Oncol.* **147** 178–85

Nenoff L *et al* 2021 Dosimetric influence of deformable image registration uncertainties on propagated structures for online daily adaptive proton therapy of lung cancer patients *Radiother. Oncol.* **159** 136–43

Nenoff L, Matter M, Charmillot M, Krier S, Uher K, Weber D C, Lomax A J and Albertini F 2021 Experimental validation of daily adaptive proton therapy *Phys. Med. Biol.* **66** 205010

Nenoff L, Matter M, Lindmar J H, Weber D C, Lomax A J and Albertini F 2019 Daily adaptive proton therapy–the key to innovative planning approaches for paranasal cancer treatments *Acta Oncol.* **58** 1423–8

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** R99

Paganetti H, Botas P, Sharp G C and Winey B 2021 Adaptive proton therapy *Phys. Med. Biol.* **66** 22TR01

Pedroni E *et al* 2004 The PSI Gantry 2: a second generation proton scanning gantry *Z. Med. Phys.* **14** 25–34

Ricchetti F, Wu B, McNutt T, Wong J, Forastiere A, Marur S, Starmer H and Sanguineti G 2011 Volumetric change of selected organs at risk during IMRT for oropharyngeal cancer *Int. J. Radiat. Oncol.***Biol.***Phys.* **80** 161–8

Scandurra D, Albertini F, van der Meer R, Meier G, Weber D C, Bolsi A and Lomax A 2016 Assessing the quality of proton PBS treatment delivery using machine log files: comprehensive analysis of clinical treatments delivered at PSI Gantry 2 *Phys. Med. Biol.* **61** 1171–81

Sharp G C, Li R, Wolfgang J, Chen G, Peroni M, Spadea M F, Mori S, Zhang J, Shackleford J and Kandasamy N 2010 Plastimatch: an open source software suite for radiotherapy image processing *Proc. 16th Int. Conf. on the Use of Computers in Radiotherapy (ICCR)*

Sherer M V, Lin D, Elguindi S, Duke S, Tan L-T, Cacicedo J, Dahele M and Gillespie E F 2021 Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review *Radiother. Oncol.* **160** 185–91

Smolders A, Lomax A, Weber D C and Albertini F 2023 Patient-specific neural networks for contour propagation in adaptive radiotherapy *Phys. Med. Biol.* **68** 095010

Sonke J-J, Aznar M and Rasch C 2019 Adaptive radiotherapy for anatomical changes *Semin. Radiat. Oncol.* **29** 245–57

Taha A A and Hanbury A 2015 Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool *BMC Med. Imaging* **15** 29

Tao C J *et al* 2015 Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study *Radiother. Oncol.* **115** 407–11

Tsuji S Y, Hwang A, Weinberg V, Yom S S, Quivey J M and Xia P 2010 Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer *Int. J. Radiat. Oncol.***Biol.***Phys.* **77** 707–14

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

van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F and Nuyts S 2019 Benefits of deep learning for delineation of organs at risk in head and neck cancer *Radiother. Oncol.* **138** 68–74

van Kranen S, Hamming-Vrieze O, Wolf A, Damen E, van Herk M and Sonke J J 2016 Head and neck margin reduction with adaptive radiation therapy: robustness of treatment plans against anatomy changes *Int. J. Radiat. Oncol., Biol., Phys.* **96** 653–60

van Rooij W, Dahele M, Ribeiro Brandao H, Delaney A R, Slotman B J and Verbakel W F 2019 Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation *Int. J. Radiat. Oncol.***Biol.***Phys.* **104** 677–84

Verhaart R F, Fortunati V, Verduijn G M, van Walsum T, Veenland J F and Paulides M M 2014 CT-based patient modeling for head and neck hyperthermia treatment planning: manual versus automatic normal-tissue-segmentation *Radiother. Oncol.* **111** 158–63

Voet P W J, Dirkx M L P, Teguh D N, Hoogeman M S, Levendag P C and Heijmen B J M 2011 Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis *Radiother. Oncol.* **98** 373–7

Wong J *et al* 2020 Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning *Radiother. Oncol.* **144** 152–8

Wu Q, Mohan R, Morris M, Lauve A and Schmidt-Ullrich R 2003 Simultaneous integrated boost intensity-modulated radiotherapy for locally advanced head-and-neck squamous cell carcinomas. I: dosimetric results *Int. J. Radiat. Oncol.***Biol.***Phys.* **56** 573–85

Yang J *et al* 2018 Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017 *Med. Phys.* **45** 4568–81

Zhang F, Wang Q, Yang A, Lu N, Jiang H, Chen D, Yu Y and Wang Y 2022 Geometric and dosimetric evaluation of the automatic delineation of organs at risk (OARs) in non-small-cell lung cancer radiotherapy based on a modified densenet deep learning network *Frontiers Oncol.* **12** (https://frontiersin.org/articles/10.3389/fonc.2022.861857)

Zhang M, Westerly D C and Mackie T R 2011 Introducing an on-line adaptive procedure for prostate image guided intensity modulate proton therapy *Phys. Med. Biol.* **56** 4947–65