Physics in Medicine & Biology

PAPER • OPEN ACCESS

Patient-specific neural networks for contour propagation in online adaptive radiotherapy

To cite this article: A Smolders *et al* 2023 *Phys. Med. Biol.* **68** 095010

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

OPEN ACCESS

RECEIVED
21 December 2022

REVISED
21 February 2023

ACCEPTED FOR PUBLICATION
5 April 2023

PUBLISHED
25 April 2023

PAPER

Patient-specific neural networks for contour propagation in online adaptive radiotherapy

A Smolders[1,2,*], **A Lomax**[1,2], **DC 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
[*] Author to whom any correspondence should be addressed.

E-mail: andreas.smolders@psi.ch

Keywords: contour propagation, adaptive radiotherapy, deep learning, biomedical image segmentation

Abstract

Objective. fast and accurate contouring of daily 3D images is a prerequisite for online adaptive radiotherapy. Current automatic techniques rely either on contour propagation with registration or deep learning (DL) based segmentation with convolutional neural networks (CNNs). Registration lacks general knowledge about the appearance of organs and traditional methods are slow. CNNs lack patient-specific details and do not leverage the known contours on the planning computed tomography (CT). This works aims to incorporate patient-specific information into CNNs to improve their segmentation accuracy. *Approach.* patient-specific information is incorporated into CNNs by retraining them solely on the planning CT. The resulting patient-specific CNNs are compared to general CNNs and rigid and deformable registration for contouring of organs-at-risk and target volumes in the thorax and head-and-neck regions. *Results.* patient-specific fine-tuning of CNNs significantly improves contour accuracy compared to standard CNNs. The method further outperforms rigid registration and a commercial DL segmentation software and yields similar contour quality as deformable registration (DIR). It is additionally 7–10 times faster than DIR. *Significance.* patient-specific CNNs are a fast and accurate contouring technique, enhancing the benefits of adaptive radiotherapy.

1. Introduction

Over the years, advanced radiation delivery paradigms such as intensity-modulated radiotherapy, volumetric modulated arc therapy and intensity-modulated proton therapy have increased the dose conformality with the tumor, resulting in improved healthy tissue sparing (Lomax 1999, Bortfeld 2006, Otto 2008, Tran *et al* 2017, Moreno *et al* 2019). However, daily set-up variations and longitudinal anatomical changes throughout the treatment, such as weight loss and tumor shrinkage, result in differences between the planned dose and the delivered dose. This can lead to target coverage degradation for highly conformal radiotherapy that may impact tumor local control. The effect is especially apparent for proton therapy, because the depth of the proton dose peak is highly dependent on the tissue densities along the beam path, which changes with changing anatomy (Lomax 2008, Zhang *et al* 2011). Uncertainties in set-up, anatomy and range are accounted for in the planning process either by applying margins around the clinical target volume (CTV) (Albertini *et al* 2011) or by incorporating the uncertainties directly using robust optimization (Liu *et al* 2012, Unkelbach *et al* 2018). However, both techniques result in an increased dose to the normal tissue, reducing the advantage of conformal radiotherapy.

With online adaptive radiotherapy, the set-up and anatomical uncertainty can be strongly reduced. The daily treatment plan is reoptimized based on a 3D daily image taken shortly before the treatment (Yan *et al* 1997, Lim-Reinders *et al* 2017, Albertini *et al* 2020, Paganetti *et al* 2021). The consequent reduction of uncertainty increases

the plan conformality and, hence, the sparing of healthy tissue. Online plan adaptation is a time and resource-intensive process as it requires the repetition of several planning steps for every fraction. In particular, it requires organs-at-risk (OARs) and target volumes delineation on the new images, plan evaluation, adaptation, reoptimization and quality assurance (QA). To be effective, all these steps need to be executed in several minutes because the time between the image acquisition and the treatment needs to be as low as reasonably possible to ensure high correspondence between the image and the treated anatomy. Furthermore, faster adaptation shortens the patient's overall treatment time and therefore increases patient comfort.

The time required for online adaptation implies automation of each sub-process with as little as possible manual interventions. The most resource-intensive step is daily contouring, so there is a great interest to automate it with sufficient accuracy and robustness (Lim-Reinders *et al* 2017). It can be automated in two distinct ways: automatic segmentation or registration.

Firstly, state-of-the-art segmentation is usually based on deep learning (DL) with convolutional neural networks (CNNs), which learn to segment medical images based on large datasets with manually annotated contours (Chen *et al* 2021, Nikolov *et al* 2021). The advantages of these methods are that they are fast, consistent and yield accurate results. On the downside, they require large amounts of annotated data to train and do not always generalize well to out-of-distribution data, e.g. scans that are significantly different than the training data. Furthermore, their applicability for tumor and target volume segmentation is limited (Kosmin *et al* 2019, Liu *et al* 2021). CNNs do not require manual contours for the patient under study, which is an advantage for segmentation in general. However, in adaptive therapy, such a reference annotation is always available, i.e. on the planning CT, containing information that is not used in the automatic segmentation of the daily scans.

Another set of methods relies on image registration for contouring (Thor *et al* 2011, Kumarasiri *et al* 2014, Elmahdy *et al* 2019). Specifically for adaptive therapy, the manual contours on the reference CT can be propagated to the daily scan by registering the former to the latter and applying the same transformation to the reference contours. The main advantage is that this technique does not require a large training dataset. The disadvantage is that it requires at least one annotated scan per patient and that traditional techniques are slow compared to auto-contouring with CNNs (Klein *et al* 2009, Costea *et al* 2022). Furthermore, when anatomical changes occur, deformable image registration (DIR) is needed, which is an ill-posed problem requiring careful hyperparameter tuning and algorithm choice to achieve high performance (Brock *et al* 2017).

To overcome the long runtime of traditional DIR algorithms, recent works have proposed image registration with deep learning (Fu *et al* 2020, Haskins *et al* 2020, Xiao *et al* 2020). Instead of iteratively optimizing a similarity metric, these CNNs are trained to directly predict the deformation which reduces the runtime strongly. However, despite the great potential, these techniques have not yet achieved the same performance as iterative algorithms (Fu *et al* 2020).

Both registration and segmentation have their advantages and disadvantages. On the one hand, iterative deformable registration is slow and can be unreliable in case of large anatomical changes or mass variations (Oh and Kim 2017, Brock *et al* 2017). On the other hand, CNNs can fail on out-of-distribution data and cannot accurately segment tumors, so they cannot be employed in adaptive therapy without time-consuming manual checks and adjustments made by clinicians. However, by including the information from the (contoured) planning CT in the CNN, its robustness can be increased because the daily images are closely related to the planning CT, so the distribution of the CNN is likely to encompass the daily images. This can be achieved by (re-) training the CNN on the planning CT, also known as patient-specific fine-tuning, which has been explored for prostate cancer on MR and CT (Elmahdy *et al* 2020, Fransson *et al* 2022), for a single OAR in the head region on CT (Chun *et al* 2021) and for brain white matter segmentation on MR (Jansen *et al* 2020).

Whereas all works report a strong improvement of the quality of the CNN by patient-specific fine-tuning, their implementation details differ and the results are specific to a single anatomical site. A rigorous comparison of this technique to other auto-contouring methods for adaptive therapy has not yet been performed, and it is therefore unclear whether it is usable and optimal.

In this work, we train patient-specific CNNs for automatic contouring in online adaptive proton therapy (PT) and compare this technique to general segmentation networks and registration-based contour propagation for patients with head and neck cancer (HNC) and non-small cell lung cancer (NSCLC). Our work differs from previous publications in:

- It uses transfer learning, as in Elmahdy *et al* (2020), Jansen *et al* (2020), but updates all parameters of the CNN instead of a subset which enhances the learning capabilities.
- It uses affine and elastic deformations along with noise addition as data augmentations to mimic the set-up and anatomical variations happening in adaptive therapy. This further prevents overfitting, making the quality of the contours less sensitive to the number of training steps during the retraining.

Table 1. Number of available ground truth labels in the CPT dataset.

OAR	Annotated scans	OAR	Annotated scans
Brainstem	306	Lacrimal gland left	251
Chiasm	308	Lacrimal gland right	251
Cochlea left	290	Lung left	106
Cochlea right	295	Lung right	108
Esophagus	116	Optic nerve left	306
Eye left	297	Optic nerve right	307
Eye right	291	Parotid left	132
Heart	103	Parotid right	136
Hippocampus left	237	Spinal cord	306
Hippocampus right	241	Thyroid	97

- The technique is evaluated for different anatomical sites. The HNC patients are representative of small anatomical changes whereas the NSCLC patients undergo larger anatomical deformation, therefore also covering a large spectrum of relevant clinical deformations in adaptive radiotherapy. Additionally, both OAR and CTV segmentation is tested.

2. Materials and methodology

This section describes the different methods for contour propagation used in this study. First, the datasets used for training and evaluation are presented. Then, the registration and segmentation-based methods are described, followed by a short description of the evaluation metrics.

2.1. Datasets

This work is based on three datasets. The first dataset is from the Center For Proton Therapy (CPT) in Switzerland and contains patients treated with proton therapy between 2013 and 2021. A total of 388 patients with various indications was included, all having at least one planning CT with or without replanning CTs, yielding a total of 464 scans with annotations. Depending on the tumor location, different OARs were contoured manually by expert medical personnel, resulting in a large variation in the number of ground truth labels for each OAR (table 1). As none of these patients underwent online adaptive therapy, this dataset is solely used to pretrain the segmentation models (see section 2.3). In the remainder of this paper, this dataset will be referred to as the *CPT dataset*.

The second dataset consists of five patients with non-small cell lung cancer (NSCLC), not included in the CPT dataset. This data has previously been described in Josipovic *et al* (2016), Nenoff *et al* (2020), Amstutz *et al* (2021), Nenoff *et al* (2021). Each patient has one planning and nine repeated voluntary deep breath hold CTs. The repeated CTs were acquired on three different days, each day consisting of three different acquisitions. However, for this study, we will consider each CT to be representative of a different fraction in online adaptive therapy. All CTs were retrospectively recontoured by expert radiation oncologists according to the clinical protocol (Nenoff *et al* 2021), which included propagating the planning contours with DIR and slice-wise manual adjustments, either in Eclipse or Velocity (Varian Medical Systems, Palo Alto, USA). This dataset will be referred to as the *NSCLC dataset*.

The last dataset consists of five patients with various indications of head and neck cancer treated with proton therapy at the CPT. Each patient has a planning CT and 4 to 7 repeated CTs acquired on separate days throughout the treatments. All patients were removed from the CPT dataset so that they were not included in pretraining the networks. Even though these patients were not treated with online adaptive therapy, the repeated CTs are representative of the daily and longitudinal anatomic and set-up variations to be expected during online adaptive therapy. The repeated CTs were retrospectively recontoured by expert radiation oncologists according to the same clinical protocol as the NSCLC scans. We will refer to this data as the *HNC dataset*.

2.2. Registration based methods

In registration-based contour propagation, the reference CT is considered the moving scan which is registered to the daily CT, i.e. the fixed scan. This registration results in a deformable vector field (DVF), which is used to interpolate the binarized reference contours to transfer them to the daily scan. In this work, we consider two distinct registration techniques.

2.2.1. Rigid registration

The first registration method relies on rigid registration (RR), i.e. the reference CT is only translated and rotated to match the daily CT. In case the anatomy is not strongly deforming (such as the head), this technique can be preferred because of its simplicity, speed and consistency. More specifically, we employ rigid registration implemented in `elastix` (Klein *et al* 2010) with mean squared error (MSE) as similarity criterion and four consecutive resolutions.

2.2.2. Deformable registration

The second registration method is a deformable image registration (DIR) method, which is the preferred method for contour propagation in case of deforming anatomy. The downside of DIR is that the problem is ill-posed, so that the results of different algorithms and even hyperparameters can lead to strongly different results (Brock *et al* 2017). In this work, we use the b-spline algorithm implemented in `plastimatch` (Sharp *et al* 2010) with MSE as similarity criterion. A detailed description of the hyperparameters can be found in (Nenoff *et al* 2021). Several other DIR algorithms were also tested, but for the sake of clarity we focus on this one as it led to good results compared to the other DIR algorithms and is publicly available.

2.3. Segmentation based methods

We train deep CNNs for the task of contour propagation in adaptive radiotherapy in two different settings: *pretrained* (or general) and *patient-specific*. All networks are based on the 3D UNet architecture, which takes as an input the daily CT and outputs a set of segmentation maps S, each map corresponding to an OAR or target volume (TV). The network has 16 initial convolutional filters, which are doubled in each of the four encoder blocks. Max pooling with kernel size and stride 2 is used for downsampling between the encoders. Four decoders upsample the features to the original resolution with nearest-neighbor interpolation. All encoders and decoders consist of 2 convolutional filters with kernel size $3 \times 3 \times 3$ followed by a rectified linear unit activation. A final convolution with kernel $1 \times 1 \times 1$ is used to convert the 16 features into a set of organ-specific activation maps. All networks are trained with binary cross-entropy, i.e the segmentation allows each voxel to be part of multiple labels. Even though organs generally do not overlap, this allows to easily handle sparsely annotated scans, i.e. scans on which some organs are visible but not segmented by the medical personnel because they were irrelevant to the planning. To still leverage all the contours in the dataset, the loss function is adjusted in such a way that it ignores the loss contributions from labels that were not manually segmented.

2.3.1. Pretrained neural network

The pretrained neural networks (PNN) are firstly trained on the relatively large CPT dataset. The models are trained from scratch with the `Adam` optimizer for 200 epochs and initial learning rate 10^{-3}, which is halved every 20 epochs. Early stopping is applied by retaining the model with the lowest loss on the validation set (10% of the patients). All scans are resampled to a fixed resolution $0.97 \times 0.97 \times 2$ mm and data augmentations include random cropping, rotations within $\pm 5°$, $\pm 5\%$ scaling, small localized elastic deformations (Isensee *et al* 2020) and Gaussian noise with $\sigma^2 = 10^{-4}$. Note that these networks do not segment any of the target volumes, because the dataset contains a wide variety of indications and previous work has shown the poor quality of CNNs for target volume segmentation (Kosmin *et al* 2019, Liu *et al* 2021).

We train two networks, one specific for the OARs in the head and neck region, i.e pretrained HNC network, and one for the OARs in the lung region, i.e pretrained NSCLC network. After training on the CPT dataset, the models are in a second step retrained on the HNC and NSCLC datasets themselves. The evaluation is done with leave-one-out validation, i.e the pretrained model is retrained on 4 out of 5 scans of either the HNC or NSCLC dataset and the retrained model is evaluated on the remaining scan. In that way, the pretrained model has still never seen the anatomy of the patient under study and should therefore generalize what it has learned from other patients. The retraining parameters are similar to the initial training parameters, but the magnitude of the data augmentations was increased to $\pm 10°$ rotations, $\pm 10\%$ scaling and $\sigma^2 = 10^{-2}$ Gaussian noise to avoid overfitting the very small dataset.

2.3.2. Patient-specific neural network

During online adaptive therapy, clinically accepted contours on the planning CT are always available because they were used for the initial planning. The pretrained networks however do not leverage these. To include this prior information, we fine-tuned the pretrained networks by retraining the networks only on the reference CT (Elmahdy *et al* 2020, Chun *et al* 2021), yielding patient-specific neural networks (PSNN). This retraining results in overfitting of the network to the reference CT, but because the reference CT is very similar to the daily CTs, it can be expected that this overfitted network still performs better than the generalizing pretrained networks. Further, to avoid complete overfitting, training is restarted with a lower learning rate 10^{-4} and, since there is

Table 2. Comparison of the average segmentation accuracy on the NSCLC OARs when fine-tuning only the last layer versus fine-tuning all weights.

Fine-tuning	Dice [%]	Surface Dice [%]	HD95 [mm]
Last layer	87.2	85.0	6.2
All layers	91.9	93.2	4.0

only one scan in the training set, runs for 50 000 epochs. Data augmentations are the same as for the initial pretraining, with the exception of a stronger Gaussian noise with $\sigma^2 = 10^{-2}$.

Pretraining a network for target volume segmentation is very difficult and would require a lot of data. However, this does not mean that the target volumes (TV) cannot be segmented with deep CNNs. Similar to the fine-tuned models, we can train a neural network solely on the reference CT, which contains the TVs contoured by a clinician. This is commonly referred to as one-shot image segmentation (Shaban *et al* 2017). Contrarily to the fine-tuned models, the training cannot restart from a pretrained neural network that is already able to segment TVs. It is however possible to leverage some prior information during one-shot learning by means of transfer learning (Weiss and Khoshgoftaar 2016), which has shown promising results in e.g video segmentation (Caelles *et al* 2017). With transfer learning, the network is first trained on a different task than it is supposed to (e.g. lung segmentation). In a second step the network is then retrained on the original task (e.g. TV segmentation) starting with the initial weights from the other training. Here, we take the pretrained models on the OARs and use transfer learning to segment the CTV. We restart the training from the final weights of the pretrained models for all layers except the final convolution, as this convolution creates organ-specific maps which are not informative for the TVs.

2.3.3. Commercial segmentation

Finally, the trained CNNs are also compared to a clinically used commercial auto-contouring software Limbus Contour 1.7 (AI Limbus Inc., 2076 Athol Street, Regina, SK S4T 3E5, Canada). This software has been clinically validated and shown to only rarely require manual adjustments of OARs (Wong *et al* 2020, D'Aviero *et al* 2022).

2.4. Evaluation methods

The performance of the above-mentioned contour propagation methods is evaluated on the HNC and NSCLC by comparing the results with the manually annotated contours on the repeat CTs. We use three well-known geometric metrics for this comparison. Firstly, the dice coefficient to evaluate the overlap between the manual and propagated contour. The dice coefficient is however strongly dependent on the size of the structure and is therefore difficult to compare for organs with different sizes. To alleviate this effect, we also include the surface dice, which represents the proportion of the organ surface which is within a tolerance of the surface of the manually annotated organ (Nikolov *et al* 2021). We set this tolerance to 2 mm. Both dice and surface dice coefficients give insight into the average difference between segmentations. To also assess the maximal error, we evaluate the 95th percentile of the Hausdorff distance (HD). A Wilcoxon signed rank test is performed between each method and the patient-specific NNs to test whether they perform significantly better or worse than the other methods.

Two preliminary experiments are performed on the NSCLC dataset to highlight the differences between the proposed method and previous works. Firstly, we compare our approach (i.e. the fine-tuning all weights of the network) to fine-tuning only the final layer, as proposed by Elmahdy *et al* (2020), Jansen *et al* (2020). Secondly, we evaluate the importance of using data augmentations, by comparing our approach to fine-tuning without data augmentations.

3. Results

3.1. Preliminary experiments

Fine-tuning all weights of the network improves the segmentation compared to only fine-tuning the last layer (table 2). This means that increasing the learning capability by retraining all weights indeed improves the performance of the network.

Including data augmentations during fine-tuning increases contouring accuracy (figure 1). For all patients, the maximum dice score during training is higher with data augmentation than without. Moreover, training with data augmentations avoids overfitting, i.e. the segmentation accuracy on the repeated CTs first increases and then stagnates, without significantly decreasing at the end of the training. Contrarily, without data augmentations, the accuracy reaches a maximum after which it steadily decreases.

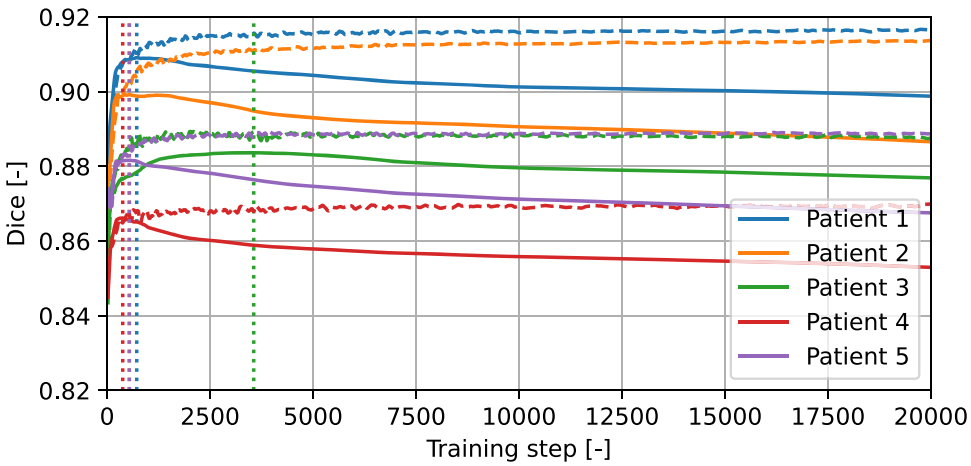

Figure 1. Evolution of the average dice score on the repeated CTs of the NSCLC dataset during fine-tuning of the PSNNs. Solid lines represent training without data augmentation, dashed lines with. The vertical dotted lines depict at which iteration the training without data augmentation reaches maximum dice for each patient.

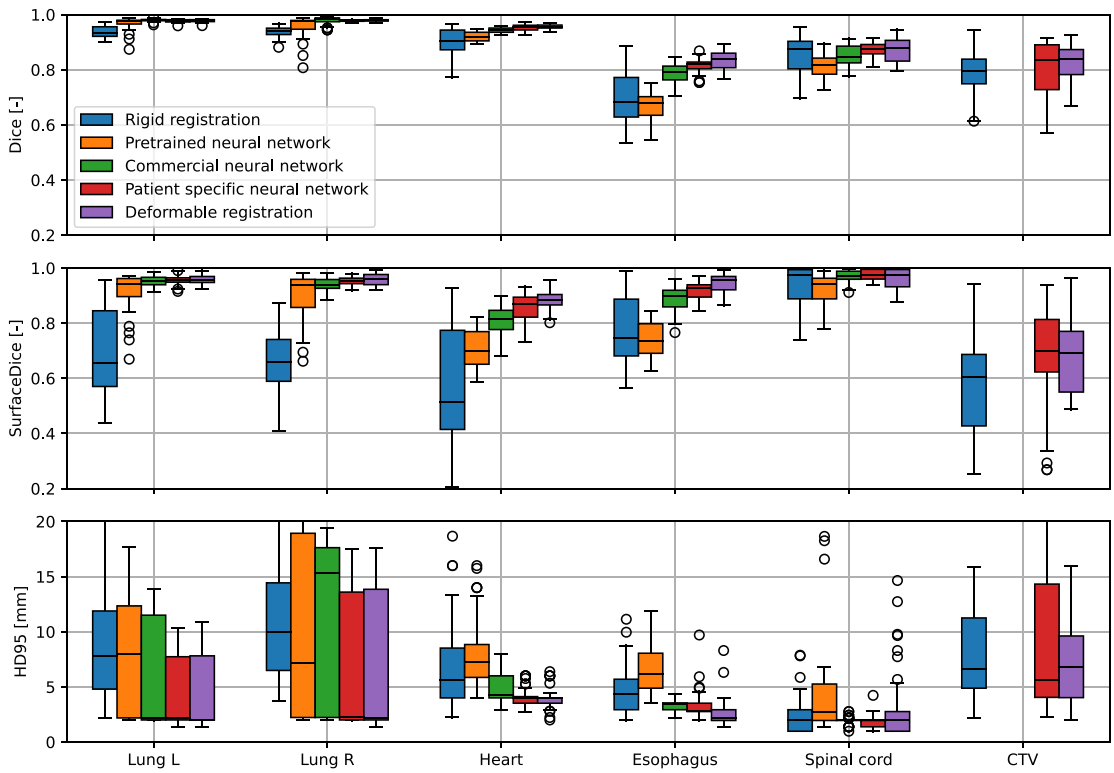

Figure 2. Comparison of the contour propagation techniques on the relevant OARs and CTV for the NSCLC patients using dice, surface dice and the 95th percentile of the Hausdorff Distance (HD95).

In practice, a fixed number of iterations needs to be defined. When training without data augmentations, the iteration at which the dice score is maximal depends on the patient (figure 1). For example, here, patient 1 reaches maximal dice after 700 iterations, and for patient 3 this is 3500. Therefore, selecting a fixed number will result in suboptimal performance for some patients. Contrarily, when training with data augmentations, the number of iterations can simply be set high (50 000 in our case) as the quality stagnates.

3.2. Contouring accuracy

Regarding the OAR contours, rigid registration (RR) performs generally worst of all methods for the NSCLC dataset (figure 2), except for the spinal cord. This is because the RR aligns the spine well, and, hence, also the spinal cord is accurately contoured. The pretrained NN achieves better contour accuracy but suffers from

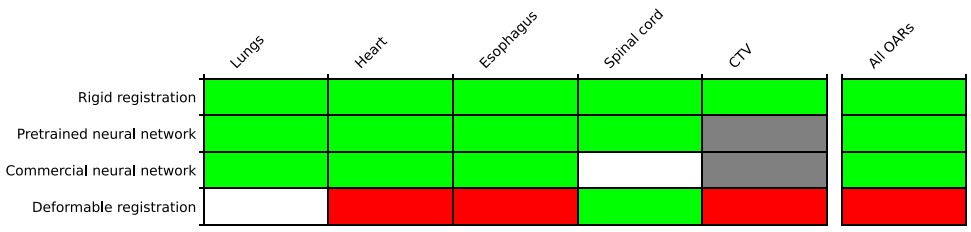

Figure 3. Overview of the Wilcoxon signed rank test results for the surface dice of the contours in the NSCLC dataset. Green: patient-specific NN performs significantly better than the method. Red: patient-specific NN performs significantly worse than the method. White: the performance of the method is not significantly different from the PSNN. Grey: the method does not segment the structure. The significance level is set to 2.5%.

outliers with low performance for the lungs and esophagus. This happens when the network is evaluated on out-of-distribution data, i.e data that is significantly different from the training data. Because the training set is small for these OARs (table 1), the probability of this is indeed larger than for the more frequently occurring OARs. The commercial system consistently outperforms the pretrained NN and is especially more robust, i.e. it suffers less from outliers. The large HD95 in the lungs in some cases is due to the presence of a tumor, which, depending on location, annotator or method is included or excluded in a contour. This has only a limited effect on the dice and surface dice, but affects strongly the HD95.

Fine-tuning the segmentation networks on a specific patient improves the segmentation accuracy of the OARs strongly, outperforming rigid registration, the pretrained NN and the commercial contouring software significantly for all OARs (figure 3). Note that we only show the significance test results for the surface dice, but similar results are found for the dice and HD95. It also resolves the outliers, because fine-tuning on the planning CT avoids that the network is run on out-of-distribution data. The contour quality is similar to DIR for the lungs, significantly lower for the heart and esophagus and significantly better for the spinal cord (figure 3).

In order to assess whether the obtained performance of the patient-specific NNs is clinically acceptable, it can be compared to the variability between the contours drawn by different observers, i.e. the inter-observer variability. This variability was not studied here, but has been quantified in other works for the relevant OARs in the thorax region (Yang *et al* 2018). It is important to note that the values are organ and image-modality-specific, as the metrics are strongly affected by the volume or contrast of the organ. The reported inter-observer dice scores were 0.96 for the lungs, 0.93 for the heart, 0.81 for the esophagus and 0.86 for the spinal cord. These values are very close to the dice scores for the patient-specific NNs and DIR (figure 2).

Regarding target segmentation, the patient-specific NNs perform significantly better than RR, but significantly worse than DIR (figure 3). However, these differences are small and only significant for the surface dice and not for dice. For one patient, the patient-specific NN has much lower contour quality. In this patient, the shape of the tumor changed throughout the treatment, causing the manual delineations to alter significantly from the reference. Whereas the performance of DIR is also low for this patient, the drop in quality is less pronounced. Such strong outliers did not occur for the patient-specific OAR segmentation, which indicates that one-shot segmentation lacks robustness because of its limited general knowledge.

Most general trends found for the NSCLC data also hold for the HNC dataset (figure 4). The main difference is that RR performs much better. For OARs in head, close to the skull (e.g. brainstem, chiasm, hippocampus), RR performs as well as the more advanced methods (figure 5), because it matches the skull accurately and the rigid assumption is applicable there. Contrarily, for OARs further from the skull (e.g. spinal cord, thyroid), RR performs badly because the rigid transformation matching the skull is not valid there. Lastly, for organs that change strongly during radiotherapy (e.g. parotid glands), RR sometimes performs very badly.

Despite the larger OAR dataset in the HN region compared to the thorax (table 1), the performance of the pretrained NN is still low. This is most apparent for the smaller OARs (e.g. lacrimal gland, optic nerve, chiasm). Again, the commercial system outperforms the pretrained NN. The patient-specific NNs significantly outperform all other methods (including DIR) for all organs in general (figure 5). However, for the individual organs, we find that the difference is not always significant and that segmentation of the optic nerves is even better with registration and the commercial segmentation.

Several other works have investigated the inter-observer variability for OARs in the head and neck region (Deeley *et al* 2011, Brouwer *et al* 2012, Mattiucci *et al* 2013, Verhaart *et al* 2014, Tao *et al* 2015, van der Veen *et al* 2019, Wong *et al* 2020). Whereas the stated values vary between the publications because of differences in experimental set-up, we found that the mean dice scores of patient-specific NNs and DIR here are similar or even higher than the reported inter-observer variabilities for all OARs except the thyroid.

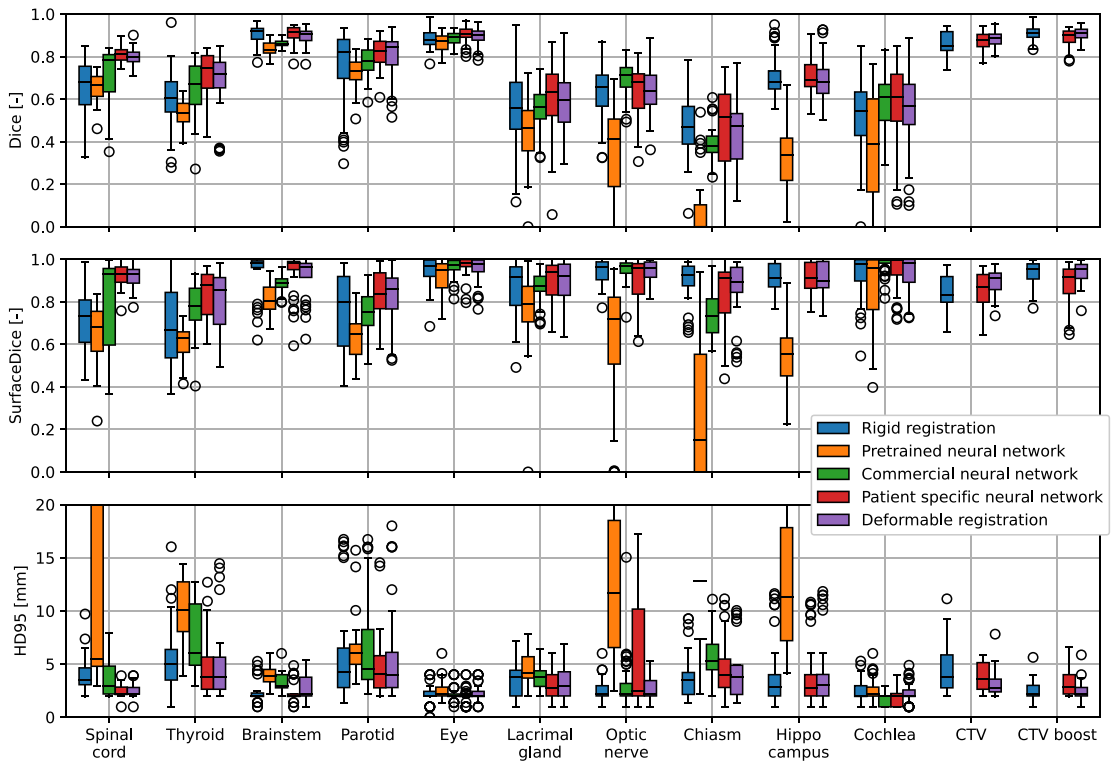

Figure 4. Comparison of the different contour propagation techniques the relevant OARs and CTV for the HNC patients using dice, surface dice and the 95th percentile of the Hausdorff Distance (HD95). The commercial segmentation did not contain hippocampus contours at the time of writing, so these metrics are not included.

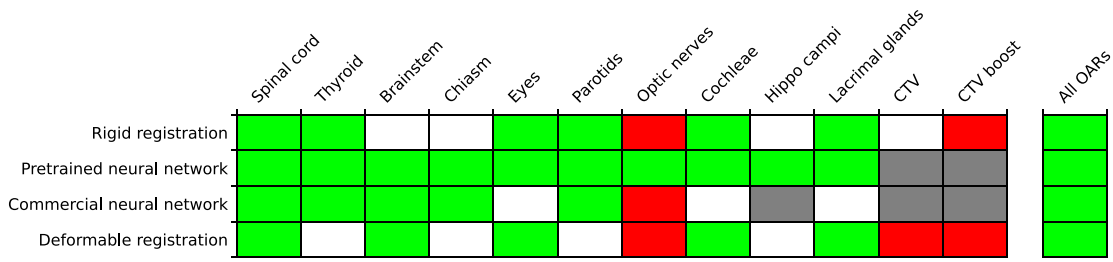

Figure 5. Overview of the Wilcoxon signed rank test results for the surface dice of the contours in the HNC dataset. Green: patient-specific NN performs significantly better than the method. Red: patient-specific NN performs significantly worse than the method. White: the performance of the method is not significantly different from the PSNN. Grey: the method does not segment the structure. The significance level is set to 2.5%.

Contouring of the main CTV works best with DIR, followed by patient-specific NNs and rigid registration. Rigid registration does not work well because the CTV covers part of the neck region, where significant shrinkage happened for these patients which cannot be captured with rigid transformations. The improvement of the patient-specific NNs compared to rigid registration is only significant based on dice score but not for the surface dice (figure 5). For the boosted region, rigid registration works well and even significantly better than the patient-specific segmentation, as this region is inside the head close to the skull.

3.3. Contouring speed

The runtime of the algorithms depends strongly on the hardware and potential GPU acceleration. Image registration in `plastimatch` and `elastix` runs on CPU and the runtime is evaluated on a Linux based system with 8 Intel Xeon E3-1240 v5 CPU cores. The runtime of the in-house trained NNs is evaluated by running inference on a Nvidia Quadro P6000 GPU and the commercial software was ran on a Nvidia RTX 3060 GPU.

Rigidly registering the CTs takes approximately the same time as running inference of the in-house trained CNNs. The commercial segmentation software is approximately 2 times slower, but still significantly faster than

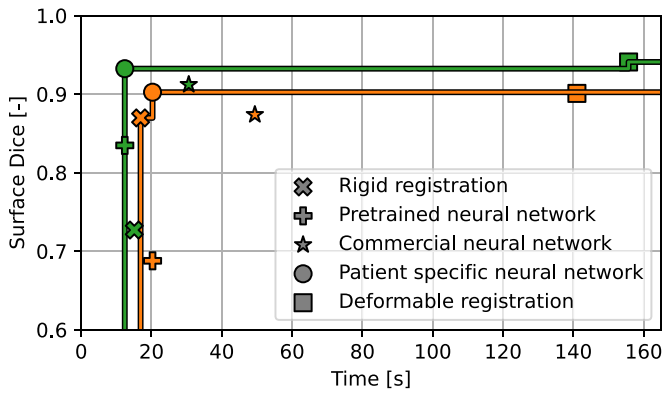

Figure 6. Mean runtime versus mean OAR surface dice for the different contouring methods and datasets. The markers correspond to the individual methods and the solid line depicts the pareto optimal front. Green: NSCLC dataset; Orange: HNC dataset.

Table 3. Average time for running contour propagation for each method and dataset.

Method	Time [s]	
	NSCLC	HNC
Rigid registration	15.1	16.9
Pretrained NN	12.4	20.3
Commercial segmentation	30.6	49.4
patient-specific NN	12.4	20.3
Deformable registration	155.7	141.0

DIR, which is 7–10 times slower than rigid registration (table 3). Note that several DIR methods with GPU acceleration exist, which could lead to significant speed up (Gu *et al* 2010, Weistrand and Svensson 2015). Whereas the runtime of the DIR is likely acceptable, the speed of the other methods offers an advantage for patient comfort and correspondence between CT and treated anatomy in a particularly time-dependent setting such as adaptive therapy. Figure 6 visualizes the trade-off between speed and accuracy. Patient-specific NNs lie on the pareto front for both HNC and NSCLC datasets, i.e. none of the other methods can improve accuracy without increasing runtime nor improve runtime without reducing accuracy. For NSCLC, also DIR lies on the pareto front, yielding slightly higher accuracy but slower runtime. For HNC, the pareto front is shared with RR, which is faster but yields lower accuracy.

4. Discussion

Our results show that DIR yields generally the most accurate contours for the targets and OARs in the thorax region. Contrarily, the patient-specific NNs are best for OARs in the head and neck region. The differences are however small and not always significant for all metrics. The PSNN is further on average 10 times faster, which is advantageous in adaptive therapy. For the HNC specifically, rigid registration is both fast and accurate for the structures close to the skull, but the accuracy is lower for those far from the skull which can lead to unacceptable degradation in target coverage.

Although both patient-specific NNs and DIR lead to high-quality contours, they do not perfectly correspond to the manual ones. This can be due to limitations of the methods, but also due to inaccuracies in the manual contours, as it has been shown that substantial inter- and inter-observer variability in delineation of HNC and NSCLC exists (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, van der Veen *et al* 2021, Kumar *et al* 2022, Zhang and Huang 2022). The PSNN and DIR methods reach accuracies similar to such inter-observer variabilities found in the literature, which impose an upper-bound for the average achievable accuracy. This further means that the methods perform similar to a human, indicating that they can be used directly in adaptive therapy.

In order to meticulously evaluate the use of contour propagation methods for adaptive therapy, the effect on the dose and the corresponding biological effect should be analyzed. Treatment plans reoptimized on automatically propagated contours should be compared to plans reoptimized on manual contours, and these

dosimetric differences have to be interpreted clinically before implementation in the clinic. This is the subject of current work at CPT.

The size of the evaluation datasets is relatively small, mainly because manual delineation of all daily CTs is a time-consuming process. For the NSCLC dataset, the clinicians completely manually recontoured because of the limited number of OARs. As the number of OARs in the HNC region is much larger, the clinicians manually adjusted contours propagated from the reference using DIR, in accordance with the current clinical protocol for replanning. Even though this creates a bias, the resulting contours are clinically acceptable and the DIR algorithm used to create these initial contours was different from the one used in this study.

The quality of the pretrained NN for the HNC dataset is low, even though the training dataset is relatively large. Especially for the smaller organs, the segmentation accuracy is largely insufficient. This could be due to the large number of OARs segmented by a single network. During training, the loss function is only slightly affected by these small organs, similar to class imbalance. This could result in the network favoring accurate segmentation of the larger structures over the smaller ones. This can be overcome by simply training one network for each structure. Even though this would lead to an increase in runtime, inference could still be parallelized or hierarchical approaches could be employed instead of splitting the image in patches (Shaheen *et al* 2021).

This analysis relies on the presence of daily CT scans, which requires an in-room CT. Although such in-room CT is present at several proton therapy centers, gantry-mounted CBCT scanners are more prevalent. In the future, also daily MRI scans might be used. The registration-based methods could easily be adjusted to allow multi-modal registration between CBCT/MRI and CT to propagate the contours. Further, a general segmentation network for CBCT/MRI could also be developed if an appropriate dataset is available. The patient-specific fine-tuning cannot be applied directly with CBCT/MRI. However,it can be applied on synthetic CTs, which are produced from the daily CBCT/MRI to reoptimize the plan in adaptive therapy. Although it is expected that the networks will work on such synthetic CTs, the quality of contours will have to be evaluated.

5. Conclusion

In this work, patient-specific CNNs were compared to general CNNs and (deformable) registration for the task of contour propagation in adaptive radiotherapy. We found that the patient-specific fine-tuning leads to higher quality contours than general segmentation networks, reaching similar quality as DIR but with a significant reduction in runtime. Fine-tuning further allows target volume segmentation, which is not yet feasible with general CNNs.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Actions under Grant Agreement No. 955956. The authors would like to thank Barbara Bachtiary, Reinhardt Krcek, Enrique Amaya and Marc Walser for contouring of daily CTs. We would further like to acknowledge Limbus AI for providing a trial version of Limbus Contour. Finally, we would like to thank Chiara Paganelli for the insightful discussions.

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.

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 *et al* 2020 Online daily adaptive proton therapy *Br. J. Radiol.* **93** 20190594

Amstutz F *et al* 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

Bortfeld T 2006 IMRT: a review and preview *Phys. Med. Biol.* **51** R363–79
Brock K K *et al* 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–e76
Brouwer C L *et al* 2012 3D Variation in delineation of head and neck organs at risk *Radiat. Oncol.* **7** 32
Caelles S *et al* 2017 One-shot video object segmentation *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*
Chen W *et al* 2021 A comparative study of auto-contouring softwares in delineation of organs at risk in lung cancer and rectal cancer *Sci. Rep.* **11** 23002
Chun J *et al* 2022 Intentional deep overfit learning (IDOL): a novel deep learning strategy for adaptive radiation therapy *Medical Physics* **49** 488– 96
Costea M *et al* 2022 Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system *Radiother. Oncol.* **177** 61 – 70
D'Aviero A *et al* 2022 Clinical validation of a deep-learning segmentation software in head and neck: an early analysis in a developing radiation oncology center *Int. J. Environ. Res. Public Health* **19** 9057
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
Elmahdy M S *et al* 2019 Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer *Med. Phys.* **46** 3329–43
Elmahdy M S *et al* 2020 Patient-specific finetuning of deep learning models for adaptive radiotherapy in prostate CT *Proc.— Int. Symp. on Biomedical Imaging 2020-April* pp 577–80
Fransson S, Tilly D and Strand R 2022 Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy *Phys. Imaging Radiat. Oncol.* **23** 38–42
Fu Y *et al* 2020 Deep learning in medical image registration: a review *Phys. Med. Biol.* **65** 20TR01
Gu X *et al* 2010 Implementation and evaluation of various demons deformable image registration algorithms on a GPU *Phys. Med. Biol.* **55** 207–19
Haskins G, Kruger U and Yan P 2020 Deep learning in medical image registration: a survey *Mach. Vis. Appl.* **31** 8
Isensee F *et al* 2020 batchgenerators—a python framework for data augmentation (https://doi.org/10.5281/ZENODO.3632567)
Jansen M J A *et al* 2020 Patient-specific fine-tuning of convolutional neural networks for follow-up lesion quantification *J. Med. Imaging* **7** 064003
Josipovic M *et al* 2016 Geometric uncertainties in voluntary deep inspiration breath hold radiotherapy for locally advanced lung cancer *Radiother. Oncol.* **118** 510–4
Klein A *et al* 2009 Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration *NeuroImage* **46** 786–802
Klein S *et al* 2010 A toolbox for intensity-based medical image registration *IEEE Trans. Med. Imaging* **29** 196–205
Kosmin M *et al* 2019 Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer *Radiother. Oncol.* **135** 130–40
Kumar S *et al* 2022 Variability of gross tumour volume delineation: MRI and CT based tumour and lymph node delineation for lung radiotherapy *Radiother. Oncol.* **167** 292–9
Kumarasiri A *et al* 2014 Deformable image registration based automatic CT-to- CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting *Med. Phys.* **41** 121712
Lim-Reinders S *et al* 2017 Online adaptive radiation therapy *Int. J. Radiat. Oncol.*Biol.*Phys.* **99** 994–1003
Liu W *et al* 2012 Robust optimization of intensity modulated proton therapy *Med. Phys.* **39** 1079–91
Liu X *et al* 2021 Review of deep learning based automatic segmentation for lung cancer radiotherapy *Front. Oncol.* **11** 717039
Lomax A 1999 Intensity modulation methods for proton radiotherapy *Phys. Med. Biol.* **44** 185–205
Lomax A J 2008 Intensity modulated proton therapy and its sensitivity to treatment uncertainties: II. The potential effects of inter-fraction and inter-field motions *Phys. Med. Biol.* **53** 1043–56
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
Moreno A C *et al* 2019 Intensity modulated proton therapy (IMPT)—the future of IMRT for head and neck cancer *Oral Oncol.* **88** 66–74
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
Nikolov S *et al* 2021 Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy *Journal of Medical Imaging Research* **23** e26151
Oh S and Kim S 2017 Deformable image registration in radiation therapy *Radiat Oncol J.* **35** 101–11
Otto K 2008 Volumetric modulated arc therapy: IMRT in a single gantry arc *Med. Phys.* **35** 310–7
Paganetti H *et al* 2021 Adaptive proton therapy *Phys. Med. Biol.* **66** 22TR01
Shaban A *et al* 2017 One-shot learning for semantic segmentation arXiv:1709.03410 *British Machine Vision Conference (BMVC)*
Shaheen E *et al* 2021 A novel deep learning system for multi-class tooth segmentation and classification on cone beam computed tomography. A validation study *J. Dentistry* **115** 103865
Sharp G C *et al* 2010 Plastimatch: an open source software suite for radiotherapy image processing *Proc. of the 16th Int. Conf. on the use of Computers in Radiotherapy (ICCR), Amsterdam, Netherlands*
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
Thor M *et al* 2011 Deformable image registration for contour propagation from CT to cone-beam CT scans in radiotherapy of prostate cancer *Acta Oncol.* **50** 918–25
Tran A *et al* 2017 Treatment planning comparison of IMPT, VMAT and 4p radiotherapy for prostate cases *Radiat. Oncol.* **12** 10
Unkelbach J *et al* 2018 Robust radiotherapy planning *Phys. Med. Biol.* **63** 22TR02
van der Veen J *et al* 2019 Benefits of deep learning for delineation of organs at risk in head and neck cancer *Radiother. Oncol.* **138** 68–74
van der Veen J *et al* 2021 Interobserver variability in organ at risk delineation in head and neck cancer *Radiat. Oncol.* **16** 120
Verhaart R F *et al* 2014 CT-based patient modeling for head and neck hyperthermia treatment planning: Manual versus automatic normal-tissuesegmentation *Radiother. Oncol.* **111** 158–63
Weiss K, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *J. Big Data* **3** 9
Weistrand O and Svensson S 2015 The ANACONDA algorithm for deformable image registration in radiotherapy *Med. Phys.* **42** 40–53

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
Xiao H, Ren G and Cai J 2020 A review on 3D deformable image registration and its application in dose warping *Radiat. Med. Prot.* **1** 171–8
Yan D *et al* 1997 Adaptive radiation therapy *Phys. Med. Biol.* **42** 123–32
Yang J *et al* 2018 Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017 *Med. Phys.* **45** 4568–81
Zhang H and Huang W 2022 Reduction of inter-observer variability using mri and ct fusion in delineating of primary tumor for radiotherapy in lung cancer with atelectasis *Int. J. Radiat. Oncol.**Biol.*Phys.* **114** e401
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