IOP Publishing Phys. Med. Biol. 68 (2023) 245027 https://doi.org/10.1088/1361-6560/ad0282

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
25 April 2023

REVISED
29 September 2023

ACCEPTED FOR PUBLICATION
11 October 2023

PUBLISHED
13 December 2023

PAPER

Deep learning based uncertainty prediction of deformable image registration for contour propagation and dose accumulation in online adaptive radiotherapy

A Smolders[1,2], **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

Keywords: deformable image registration, contour propagation, dose accumulation, adaptive radiotherapy, proton therapy

Abstract

Objective. Online adaptive radiotherapy aims to fully leverage the advantages of highly conformal therapy by reducing anatomical and set-up uncertainty, thereby alleviating the need for robust treatments. This requires extensive automation, among which is the use of deformable image registration (DIR) for contour propagation and dose accumulation. However, inconsistencies in DIR solutions between different algorithms have caused distrust, hampering its direct clinical use. This work aims to enable the clinical use of DIR by developing deep learning methods to predict DIR uncertainty and propagating it into clinically usable metrics. *Approach.* Supervised and unsupervised neural networks were trained to predict the Gaussian uncertainty of a given deformable vector field (DVF). Since both methods rely on different assumptions, their predictions differ and were further merged into a combined model. The resulting normally distributed DVFs can be directly sampled to propagate the uncertainty into contour and accumulated dose uncertainty. *Main results.* The unsupervised and combined models can accurately predict the uncertainty in the manually annotated landmarks on the DIRLAB dataset. Furthermore, for 5 patients with lung cancer, the propagation of the predicted DVF uncertainty into contour uncertainty yielded for both methods an *expected calibration error* of less than 3%. Additionally, the *probabilisticly accumulated dose volume histograms* (DVH) encompass well the accumulated proton therapy doses using 5 different DIR algorithms. It was additionally shown that the unsupervised model can be used for different DIR algorithms without the need for retraining. *Significance.* Our work presents first-of-a-kind deep learning methods to predict the uncertainty of the DIR process. The methods are fast, yield high-quality uncertainty estimates and are useable for different algorithms and applications. This allows clinics to use DIR uncertainty in their workflows without the need to change their DIR implementation.

1. Introduction

Intensity modulated photon and proton therapy, as well as volumetric modulated arc therapy, offer high conformality of the dose to the tumor and therefore spare healthy tissue more than traditional radiotherapy techniques (Lomax 1999, Bortfeld 2006, Otto 2008, Tran *et al* 2017, Moreno *et al* 2019). However, the effectiveness of these techniques depends on accurate dose delivery, which can be compromised by changes in the patient's anatomy or set-up variations. Such variations are especially important for proton therapy, where the location of the dose peak is highly sensitive to tissue densities along the beam path (Lomax 2008, Zhang *et al* 2011). To compensate for these uncertainties, treatment plans are often made more robust by adding margins around the target area (Albertini *et al* 2011) or using robust optimization techniques (Liu *et al* 2012, Unkelbach *et al* 2018). Whereas these

approaches can help to ensure sufficient target coverage, they also increase the dose to healthy tissue and therefore diminish some of the potential benefits of conformal therapy.

Instead of accounting for anatomical and set-up uncertainty, adaptive radiotherapy aims to reduce it by acquiring a 3D image shortly before the treatment. Reoptimizing the treatment plan based on this daily information reduces the need for robustness, and, hence, lowers the dose to the healthy tissue. To minimize the uncertainty about the treated anatomy, the time between image acquisition and treatment needs to be as short as reasonably possible, in the order of minutes. This excludes manual execution of many adaptation steps, so adaptive therapy requires extensive automation.

One tool enabling adaptive therapy is deformable image registration (DIR), i.e. finding a (deformable) transformation that maps one image to another. DIR is used for three distinct applications in adaptive therapy: contour propagation, dose accumulation and synthetic CT generation (Rigaud *et al* 2019, Paganetti *et al* 2021). For contour propagation, the planning CT is registered to the daily image and the resulting transformation is applied to the planning contours of the organs-at-risk (OARs) and target volumes (TVs). This results in contours on the daily image, which are necessary for reoptimization of the treatment plan. As this needs to happen between image acquisition and treatment, it should be relatively fast, i.e. in the order of minutes. Several works have shown great potential for DIR for contour propagation, both geometrically and dosimetrically (Hardcastle *et al* 2013, Kumarasiri *et al* 2014, Smolders *et al* 2023a, 2023b).

Another application of DIR is dose accumulation, i.e. the process of summing up doses from different time points on a common reference anatomy (Murr *et al* 2023). Dose accumulation has many applications outside adaptive therapy, e.g. to correctly evaluate the cumulative delivered dose in case of re-irradiation. In adaptive therapy, it is even more important. Due to the daily reoptimization, the treatment plan varies from day to day. To ensure that the total treatment adheres to the prescribed dose, all the daily doses have to be accumulated. Contrary to contour propagation, dose accumulation does not need to happen during the patient appointment. It could be part of an offline quality assurance (QA) procedure, e.g. after the patient treatment or once per week, so the time constraint is less stringent.

Whereas DIR has a large potential, its clinical use is currently limited. Due to the technical limitations of medical imaging (such as resolution and contrast), the corresponding points on two images of the same anatomy cannot be found in an unambiguous manner. This requires practical DIR algorithms to impose additional constraints (e.g. smoothness) in the optimization. Since the implementation and strength of these constraints vary between DIR algorithms, their solutions usually differ (Brock *et al* 2017).

To overcome the distrust in DIR and enable its use in the clinic, several works have developed methods to predict the uncertainty of a DIR solution. Some authors have proposed probabilistic DIR algorithms that directly output a distribution of solutions rather than a unique solution (Simpson *et al* 2013, Heinrich *et al* 2016). The downside of these models is that they are often relatively slow, that the quality of the solutions does not reach those of non-probabilistic algorithms or that their implementations are complex, so they are rarely used in practice. Otherwise, Monte-Carlo (MC) sampling could be used to quantify DIR uncertainty. By sampling the uncertain input parameters, such as regularization strength, or sampling from different DIR algorithms, the spread in DIR solutions can be quantified. However, such sampling would require many independent DIR runs, which, given the long DIR runtimes, would be inadequate for adaptive therapy. Another work tried to predict the uncertainty of a given DIR solution using linear regression (Amstutz *et al* 2021). Whereas this method is fast and has the advantage of being independent of the DIR algorithm itself, the uncertainty prediction is solely based on the magnitude of the transformation and disregards any image information.

In our work, we develop deep learning based uncertainty prediction methods for DIR. This has, to the best of our knowledge, not yet been attempted. The methods do not calculate the registration, but aim to predict the uncertainty of a given DIR solution. We focus on CT to CT registration, therefore either assuming that the daily image is a CT or that a daily synthetic CT is available. A supervised, unsupervised and combined model are trained and their results are compared for both contour propagation and dose accumulation. The network architecture, loss functions, datasets and training parameters are detailed in section 2. Section 3 discusses the model calibration and compares the performance for the respective applications. These results are followed by a discussion in section 4 and a general conclusion in section 5.

2. Materials and methodology

2.1. Gaussian deformable vector field (DVF)

Our work aims to predict the uncertainty of a given DVF, the output of a DIR algorithm. The DVF together with its uncertainty can be considered as a probabilistic DVF z. Whereas z can in principle have any distribution, it is here assumed to be a multivariate Gaussian, i.e.

$$z \sim q(z) = \mathcal{N}(\mu, \Sigma) \tag{1}$$

with μ the mean DVF given by an existing DIR algorithm and Σ the covariance matrix. In 3D, μ contains $3 \times H \times W \times D$ elements, with H, W, D the dimensions of the image. Our methods aim to predict Σ, which has $(3 \times H \times W \times D)^2$ elements, containing not only the variance of each vector, but also its covariance with other vectors. For an average CT of size $512 \times 512 \times 100$, storing Σ in float precision would require 24 petabytes, which is impossible for practical applications.

Therefore, we further assume a fixed correlation matrix ρ between neighboring voxels, i.e.

$$\Sigma = G\rho G^T \tag{2}$$

with G a diagonal matrix with its elements representing the standard deviations $\sigma_{p,x}$, $\sigma_{p,y}$ and $\sigma_{p,z}$ of each component of each vector of the DVF. The models only predict G, which has $3 \times H \times W \times D$ nonzero elements, requiring only 314 MB of storage for an average CT. Lastly, we define ρ as

$$\rho = C_{\sigma_c} C_{\sigma_c}^T \tag{3}$$

with C_{σ_c} the matrix corresponding to Gaussian smoothing with kernel width σ_c, which correlates vectors that are spatially close (Dalca *et al* 2019). The larger σ_c, the larger the correlation between the vectors which results in smoother DVF samples. With these assumptions, DVF samples z_k can be generated as

$$z_k = \mu + GC_{\sigma_c}\epsilon_k \tag{4}$$

with $\epsilon \sim \mathcal{N}(0,\ 1)$, a tensor of uncorrelated standard normally distributed elements (Kingma *et al* 2015). The matrix product $C_{\sigma_c}\epsilon$ does not need to be calculated explicitly: after reshaping ϵ into the DVF shape $(3 \times H \times W \times D)$, it is simply a Gaussian smoothing of ϵ correlating the neighboring elements (Dalca *et al* 2019). Note that the elements of C_{σ_c} need to be rescaled with the sum of the squared elements of the smoothing kernel, so that the individual elements of $C_{\sigma_c}\epsilon$ remain standard normally distributed and only correlated with their neighbors.

2.2. DIR algorithms

The neural networks predicting G were trained using `plastimatch`b-spline (Sharp *et al* 2010), but other DIR algorithms were included in the evaluation to verify whether the approach can be generalized to other DIR algorithms without retraining. The included DIR algorithms are shortly detailed hereafter:

- **Plastimatch b-spline**: many (commercial) DIR algorithms use a b-spline approach because it is fast and generally yields good results (Rueckert *et al* 1999, Sotiras *et al* 2013). Our models were trained using the `plastimatch` b-spline algorithm because it is scriptable and publicly available. Recent work further showed good results for contour propagation for OAR contours in both head and neck and lung cancer patients (Smolders *et al* 2023b). Another advantage is that the hyperparameters can be changed, which is necessary for our supervised training (see further). Unless stated otherwise, the optimization used mean-squared-error (MSE) as similarity criterion, two consecutive resolutions with grid spacing s respectively 40 and 20 mm and regularization $\lambda_r = 0.002$ (Nenoff *et al* 2020).
- **Plastimatch demons**: demons is a commonly used DIR algorithm (Thirion 1998). Contrary to the b-spline algorithm, it is a non-parametric method. Here, `plastimatch` demons was used. The implementation details can be found in the supplementary material of Nenoff *et al* (2020).
- **Velocity**: the Velocity DIR algorithm (Varian Medical Systems, Palo Alto, USA) was further included as it is clinically used DIR at the Center for Proton Therapy (CPT).

Three additional commercial DIR algorithms were used to benchmark dose accumulation uncertainty. These include Raystation Anaconda (RaySearch Laboratories AB, Stockholm, Sweden), Mirada (Mirada Medical, Oxford, UK), and a preclinical DIR algorithm provided by Cosylab (Cosylab, d. d., Control System Laboratory, Ljubljana, Slovenia).

2.3. Datasets

This work uses 4 different datasets (table 1). The first dataset contains CT scan pairs of patients treated at CPT between 2013 and 2021 and was used to train the neural networks. A total of 50 patients were included, which, because some had more than one repeated CT, yielded a total of 63 scan pairs. Since each scan of a pair can be considered once as the fixed and once as the moving image, this results in 126 input instances. All scan pairs were initially rigidly registered using the `Elastix` toolbox (Klein *et al* 2010) and resampled to a fixed resolution $1.95 \times 1.95 \times 2$ mm.

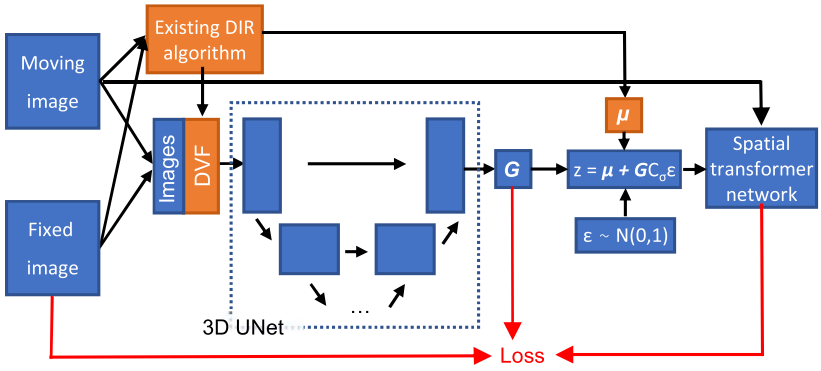

Figure 1. Schematic network architecture.

Table 1. Overview of the datasets.

Name	No. patients	No. scans	Anatomical region
CPT dataset	50	106	head and neck (64%), thorax (14%), abdomen (22%)
DIRLAB dataset (Castillo *et al* 2009, 2010)	10	20	thorax
NSCLC dataset	5	50	thorax
HNC dataset	5	33	head and neck

The performance of the models was quantitatively assessed using the public DIRLAB dataset (Castillo *et al* 2009, 2010). It contains 10 patients with in- and exhale CT scans with each 300 manually annotated landmarks (LMs). 4 patients were used for hyperparameter tuning, and 6 for evaluation.

The last two datasets contain each 5 patients with respectively non-small cell lung cancer (NSCLC) and various indications of head and neck cancer (HNC). None of them were included in the training dataset. This data has been described in Smolders *et al* (2023b). All patients each have one planning CT and several repeated CTs, all with manually contoured OARs and clinical target volumes (CTV). The NSCLC patients each had 9 repeated CTs and OARs including lungs, heart, spinal cord and esophagus. The HNC patients had 4–7 repeated CTs, and the OARs included in this work consist of the eyes, optic nerves, chiasm, brainstem, parotids, spinal cord and thyroid. Even though these patients were not treated with adaptive therapy, the repeated scans can be considered as example daily CTs in an adaptive treatment.

For the NSCLC and HNC datasets, an adaptive proton therapy treatment was simulated by reoptimizing the treatment plan on each repeated CT. Only the spot positions and weights were reoptimized, and the field angles and OAR constraints were kept the same. The NSCLC plans delivered a 2 Gy RBE fraction dose with three fields with angles and constraints specific for each patient to maximize organ sparing. The HNC plans delivered 1.6 Gy RBE with a 2.2 Gy RBE to a boosted region and all had the same 3-field configuration with gantry angles 65°, 180° and 295°. OAR constraints were imposed in line with standard clinical objectives and were the same for all patients. This yields for each patient a series of different daily dose distributions, which can be accumulated on a reference CT to validate the entire treatment.

2.4. Model architecture

The deep learning models predicting G are based on a 3D UNet architecture (figure 1) (Çiçek *et al* 2016). The fixed and moving images are given as input to an existing DIR algorithm and the resulting DVF is considered as the mean field μ. This DVF is concatenated to both images, yielding a $5 \times H \times W \times D$ tensor as network input (appendix A). Based on this input, which contains information about the magnitude of the deformation and the local image contrast, the network aims to predict a $3 \times H \times W \times D$ tensor G, containing the voxel-wise standard deviations $\sigma_{p,x}$, $\sigma_{p,y}$ and $\sigma_{p,z}$ of the DVF. G is in part directly used in the loss function. For unsupervised training, G is additionally used together with μ to generate DVF samples (equation (4)), which are used to warp the moving image in the *spatial transformer network* (Jaderberg *et al* 2015). These sampled moved images are also used in the loss function to compare to the fixed image.

The 3D UNet has an initial convolution creating 16 feature maps, which are doubled in each of the 3 consecutive encoder blocks. Each encoder and decoder block consists of 2 convolutional layers with kernel 3 and stride 1, each followed by a ReLU activation function. Downsampling between the blocks is done with max

pooling with kernel 2 and stride 2. Upsampling uses nearest-neighbour interpolation. A final convolution with kernel $1 \times 1 \times 1$ is used to convert the 16 features into G.

2.5. Loss function

The above network architecture is used in three ways, with the only difference being the loss function used during training: supervised, unsupervised and a combination of both.

2.5.1. Supervised learning

The `plastimatch` b-spline algorithm has two important hyperparameters: the grid spacing s, i.e. the distance between the b-spline control points, and the regularization λ_r, a hyperparameter in the objective function regulating the importance of the smoothness of the DVF. Depending on the patient, anatomical location and magnitude of the deformation, different hyperparameters are optimal. Instead of using a single value for λ_r and s, we can assume probability distributions $p(s)$ and $p(\lambda_r)$, and this uncertainty in hyperparameters can be propagated using MC sampling to the corresponding DVF uncertainty (Smolders *et al* 2022b). Even though this is too slow in practice, it can be used to generate standard deviation labels G_{gt}, which can be used to train a neural network.

For each image in the training set, $p(s)$ and $p(\lambda_r)$ were sampled 100 times and the b-spline algorithm was run with the sampled hyperparameters, yielding 100 DVF samples. G_{gt} was then calculated as the standard deviation of these samples with respect to the mean DVF μ

$$G_{gt} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(z_k - \mu)^2} \tag{5}$$

with $N = 100$ the number of samples. Note that μ is not calculated as the mean over these 100 samples, but as the algorithm's output when run with hyperparameters μ_s and μ_{λ_r}, as would be the case upon inference. The network is then trained with the mean absolute error (MAE), i.e.

$$\mathcal{L}_s(\Psi, f, m, \mu) = |G - G_{gt}|. \tag{6}$$

with f and m respectively the fixed and moving images, μ the mean DVF and Ψ the network parameters.

To sample $p(s)$ and $p(\lambda_r)$, their parameters need to be specified. Since grid spacings below the voxel spacing or above the image size are pointless, s is naturally bound and $p(s)$ is therefore assumed uniform. Because $\lambda_r \geqslant 0$, but is not bound in magnitude, we assumed a lognormal distribution. Even though s is naturally bound, very high or low values likely result in inaccurate DVFs for all practical applications, and should therefore not be considered. To find reasonable parameters of both distributions, the b-spline algorithm was run for a range of parameters s and λ_r on the DIRLAB validation scans, and for each hyperparameter set, the target registration error (TRE) was evaluated. Hyperparameters for which the TRE is high are less likely to be the optimal ones, so the parameters of the distributions were chosen based on this TRE.

2.5.2. Unsupervised learning

Following Dalca *et al* (2019), the unsupervised model is trained based on a variational Bayes method. With z the DVF, the network aims to minimize the Kullback–Leibler (KL) divergence between the posterior probability $p(z|f, m)$ and the Gaussian DVF $q(z)$. It is additionally assumed that:

- f is a noisy observation of the transformed moving image $m \circ z$: $p(f|m, z) = \mathcal{N}(m \circ z, \sigma_I^2 I)$;
- the prior probability is given by $p(z) = \mathcal{N}(0, \Sigma_z)$, with $\Sigma_z^{-1} = \Lambda_z = \lambda(D - A)$, λ a hyperparameter, D the graph degree matrix and A the adjacency matrix;
- $q(z) = \mathcal{N}(\mu, G\rho G^T)$ with $\rho = C_{\sigma_c} C_{\sigma_c}^T$, as indicated above.

Minimizing the KL divergence with the above assumptions yields a loss function (Dalca *et al* 2019, Smolders *et al* 2022a):

$$\begin{aligned}\mathcal{L}_u(\Psi, f, m, \mu) = &\frac{1}{2\sigma_I^2 K}\sum_k ||f - m \circ z_k||^2 \\ &+ \sum_{i=1}^{m}\sum_{j \in N(i)}\left(\frac{\lambda}{2}(D - A)_{i,j}\rho_{i,j}G_{i,i}G_{j,j}\right) \\ &- \frac{1}{2}\log(|GG^T|) + cte\end{aligned} \tag{7}$$

with K the number of samples (equation (4)), i, j respectively the rows and columns of Σ, i.e. indexing the voxels, and $N(i)$ the neighbouring voxels of voxel i, including itself. By construction, $(D - A)_{i,j} = 6$ if $i = j$ and $(D - A)_{i,j} = -1$ if $i \neq j$. Because of the fixed ρ, the components $\rho_{i,j}$ have only to be computed once at the beginning of the training. K was set to 1 to limit GPU memory usage.

The unsupervised loss function contains two hyperparameters: λ and σ_I^2. Their value influences both the magnitude of the predicted uncertainty as well as the trade-off between contrast and uncertainty. To tune these hyperparameters, the probability of observing the moving landmarks $\vec{x}_m$ given the predicted probabilistic DVF was maximized for the DIRLAB validation scans. This can be calculated as

$$p(LMs) = \prod_{i}^{CTs} \prod_{j}^{LM} p(\vec{x}_{m,i,j}|DVF_i), \tag{8}$$

assuming that each landmark is independent of the others, which is reasonable if the landmarks are sufficiently far apart. Because the Gaussian distribution is continuous, the probability of observing exactly $\vec{x}_m$ is infinitesimally small. We therefore maximized the probability that $\vec{x}_m$ is observed within a cube of 1 mm^3 around it with a homogeneous probability density, which is equal to the probability density at $\vec{x}_m$. The 1% points with the lowest probability were further discarded, because $p(LMs)$ is heavily affected by these outliers. The mean $\log p(LMs)$ is reported, making the metric independent of the number of landmarks.

The width σ_c of the Gaussian smoothing during sampling is not learned but taken as a constant value for the whole image and dataset. However, its value needs to be tuned: an overly large value would result in overly smooth DVF samples, and a too-small value could lead to folding voxels. Furthermore, the larger σ_c, the larger the width of the smoothing kernel has to be, increasing the sampling time. σ_c was therefore iteratively increased until the average fraction of folding voxels on the DIRLAB data was below 0.01% for the model with optimal λ and σ_I^2. The final values were $\sigma_c = 15$ voxels and kernel width 61, which are used for all models in this work.

2.5.3. Combining super- and unsupervised models

The super- and unsupervised methods are based on different assumptions and quantify the uncertainty differently. The one is likely more accurate in some cases and anatomical regions, and the other one in other cases. As it is *a priori* difficult to predict which method would be most appropriate for each case, a conservative approach is to predict the maximum uncertainty of both methods.

In principle it should be possible to change the loss function during training to predict this maximum directly. However, we found that the resulting discontinuity in the loss around the maximum impeded smooth convergence, even by setting different weighting factors for both loss terms We therefore opted to combine the results of both the super- and unsupervised in postprocessing by taking the maximum σ_p for each individual voxel. This technique is referred to as the *combined model*. Even though this doubles the inference time of the neural network, the total runtime of the applications (see later) is affected much less since inference accounts for only a minor fraction of the total runtime.

2.6. Training

All models were implemented in `Pytorch` and trained for 250 epochs with an initial learning rate $5e-4$, which was halved every 50 epochs. `Adam` was used for optimization (Kingma and Ba 2014). All models used random cropping as data augmentation, with patch size $256 \times 256 \times 96$, and the unsupervised model additionally used axis-aligned flipping. The models were trained using NVIDIA RTX 2080 Ti GPUs with 12 GB VRAM.

The training of the unsupervised network sometimes failed and diverged because of the sampling process used to calculate the loss function. To limit GPU memory, we only used one sample to calculate the loss function. If this sample is at the tails of the distribution, this might cause a very high loss and hence sudden divergence of the training, especially if this happens early during training when the predicted uncertainties are not necessarily reasonable. Using more samples would avoid this, but requires more GPU memory. Instead, for the hyperparameter tuning, model training was restarted using, as initial weights, the parameters of the initially converged unsupervised network, which simplified convergence as the predicted uncertainties are already in a reasonable range.

2.7. Evaluation

Since a complete ground truth deformation does not exist for patient images (Brock *et al* 2017), DIR accuracy is, for clinical purposes, often evaluated using manually annotated landmarks (Brock *et al* 2005, Weistrand and Svensson 2015, Kadoya *et al* 2016, de Vos *et al* 2019, Hering *et al* 2023) or contours (Klein *et al* 2009, Weistrand and Svensson 2015, Balakrishnan *et al* 2019, de Vos *et al* 2019, Hering *et al* 2023). Similarly, DIR uncertainty can also be evaluated using landmarks and contours. Here, the DVF uncertainty was directly evaluated using the landmarks on the DIRLAB test scans. The uncertainty prediction was additionally evaluated for both contour propagation and dose accumulation. For both applications, it is first explained how DVF uncertainty can be

propagated into contour or dose uncertainty, after which the evaluation methods are explained. The runtime of the applications was evaluated using an NVIDIA RTX 2080 Ti GPU with 12 GB VRAM.

2.7.1. Landmark uncertainty

The quality of the uncertainty prediction was assessed by splitting the predicted standard deviations σ_p into bins, and grouping all landmarks within each bin together, similar to a reliability diagram (Guo *et al* 2017). For an infinite number of landmarks and a Gaussian distribution of DVFs, the root-mean-squared error (RMSE) of the landmarks within each group should be equal to the mean σ_p of that group. We compared σ_p to the RMSE to assess the quality of the uncertainty prediction and to compare different models. Analogous to the *expected calibration error* (see further and Guo *et al* 2017), we define the *landmark expected calibration error (LECE)* as

$$\text{LECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{RMSE}(B_m) - \sigma_p(\bar{B}_m)| \tag{9}$$

with B_m a bin of landmarks, M the number of bins, $|B_m|$ the number of landmarks in the bin, n the total number of landmarks, $RMSE(B_m)$ the root-mean-squared error of the landmarks and $\overline{\sigma_p(B_m)}$ the average predicted uncertainty in the bin. In this work, the width of each bin is set to 0.25 mm.

2.7.2. Contour propagation

Contours can be propagated by sampling the DVF distribution (equation (4)) and warping the contours with the DVF sample. By doing this several times, like MC sampling, a set of potential contour samples are obtained on the daily CT, hereafter referred to as *probabilistic contour propagation*. For each voxel, the proportion of samples for which it lies within the contour sample yields an estimate of the confidence that this voxel is part of the contour, i.e. *the contour uncertainty*. For example, if a voxel lies within 65 out of 100 contour samples, the confidence that this voxel is part of the corresponding structure is considered 65%.

To evaluate the accuracy of these confidences, the planning contours of the NSCLC and HNC patients were probabilistically propagated to each repeated CT, using 100 samples. For each voxel in every repeated CT, this results in one confidence per propagated contour (OARs and CTV). Similar to the evaluation of σ_p, the predicted confidences can be split up into intervals, and the voxels with confidence within each interval can be grouped together. Using the manual contours, the proportion of voxels of each group that were inside the respective contour can be calculated. For a well-calibrated model, the average confidence of each group should be equal to this proportion. We evaluated our trained models for contour propagation by comparing the predicted confidence to the target proportions using a *reliability diagram*, as well as by calculating the *expected calibration error (ECE)* (Guo *et al* 2017). Only voxels with confidences between 1% and 99% are considered in the ECE, because otherwise it would be artificially low due to the high number of voxels very far from the contour, for which it is easy to predict 0% confidence.

2.7.3. Dose accumulation

Similar to contour propagation, the DVF distributions can be sampled to obtain an accumulated dose sample. However, several DVFs need to be sampled to obtain one accumulated dose sample (figure 2). First, each dose on the repeated CTs is warped with a sample of its corresponding DVF to the planning CT. This means that n different probabilistic DVFs are each sampled one time. Secondly, these warped doses on the planning CT are accumulated, yielding one accumulated dose sample. Repeating this K times yields K accumulated dose samples, which allows to calculate the accumulated dose uncertainty. We refer to this process as *probabilistic dose accumulation*.

As each accumulated dose sample requires one DVF sample for each repeated CT, acquired at different times through the treatment course, additional assumptions need to be taken about the temporal correlation between these DVF samples. This correlation is highly complex and depends on the patient, anatomical location and DIR algorithm. For simplicity, only two cases are considered: *fully correlated* (FC) and *not correlated* (NC). In the case of full correlation, the same ϵ_k is used for each DVF sample of the different CTs (equation (4)). Without correlation, ϵ_k is sampled independently for each DVF.

The doses of the repeated CTs were probabilistically accumulated on the planning CT with 50 samples for each NSCLC and HNC patient, once *fully correlated* and once *not correlated*. For each accumulated dose sample, a dose volume histogram (DVH) was created. These DVHs were combined in a *probabilistic DVH* , with the lower and upper bound of the DVH representing for each volume increment the 2.5th and 97.5th percentile of all sampled doses.

Whereas manual contours allow the evaluation of contour uncertainty, evaluating the dose accumulation uncertainty is difficult because the ground-truth accumulated dose is unknown. However, the uncertainty can still be evaluated by considering how well it matches a set of plausible deformation samples. In previous work (Nenoff *et al* 2020, Amstutz *et al* 2021, Smolders *et al* 2023c), the solutions of several independent DIR

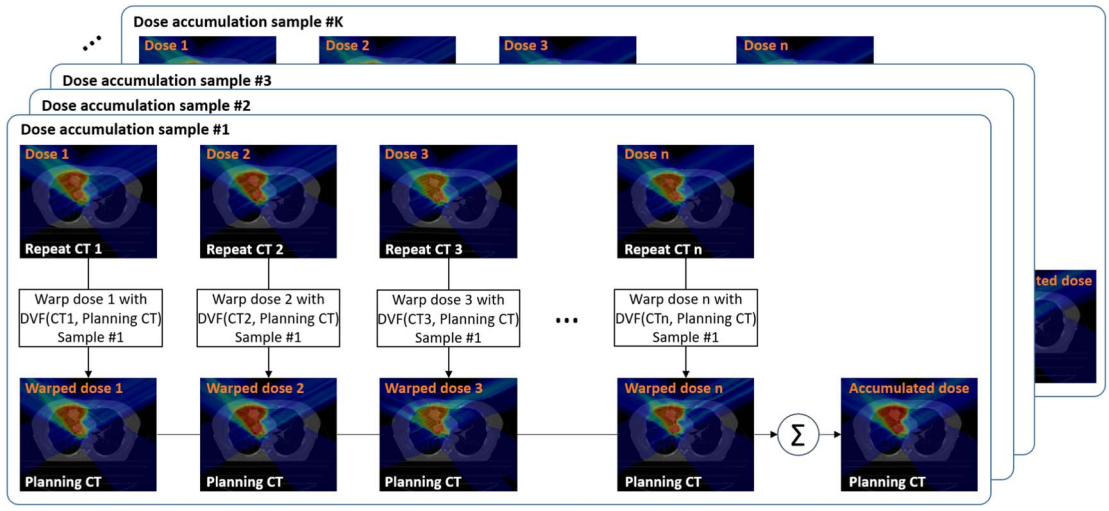

Figure 2. Schematic overview of probabilistic dose accumulation.

algorithms were assumed to be plausible deformations and were used to quantify DIR and accumulated dose uncertainty. Similarly, we accumulated the dose with 6 different DIR algorithms (section 2.2) and evaluated the volume fraction for which the DVHs of these accumulated doses lie within the error bounds of the *probabilistic DVH* , hereafter referred to as the *encompassed volume fraction EVF*. An accurate EVF, i.e. close to 95%, therefore means that our prediction matches well the uncertainty resulting from accumulating with different DIR algorithms. For this evaluation, a threshold of 0.1% was added to both bounds. This is necessary because homogeneous dose regions cause both bounds and the DVHs of other DIR algorithms to lie very close to each other. In that case, interpolation effects can cause the DVHs to lie outside the bounds, but the difference is clinically insignificant, and should therefore not be reflected in the metrics.

3. Results

3.1. Hyperparameter tuning

The procedure to tune the hyperparameters of the supervised and unsupervised models can be found in appendix B.

3.2. Model comparison

The magnitude and spatial distribution of the predicted uncertainty of the unsupervised, supervised and combined models is different (figure 3). This is due to the different assumptions on which the models are based. The supervised model predicts high uncertainty in regions with large deformation or non-correspondences, as the DIR solutions of the b-spline model vary strongly with the hyperparameters in those regions. The unsupervised model predicts high uncertainty in regions with low contrast and inversely, and is not affected by the magnitude of the deformation. In the following, the models are quantitatively compared for landmark prediction, contour propagation and dose accumulation in combination with the b-spline algorithm. For the unsupervised model, the results using other DIR algorithms are also included.

3.2.1. Landmark uncertainty prediction

For all models, there is a positive correlation between the predicted uncertainty σ_p and the RMSE, i.e. the errors are on average larger for landmarks with higher uncertainty (figure 4). For the supervised model, the landmark RMSE is systematically above the predicted uncertainty σ_p, meaning that this model underestimates the uncertainty in the DIRLAB landmarks (figure 4(a)). This is because the uncertainty in hyperparameters does not cover the total DIR uncertainty. The LECE of the unsupervised and combined models is 0.65 and 0.64 mm respectively, much lower than the LECE = 1.34 mm of the supervised model. That means that their uncertainty prediction is more accurate. In case of large uncertainty, both models underestimate the uncertainty (RMSE $> \sigma_p$ for the right part of figure 4(a)), but this is only for a limited number of landmarks. The difference between the unsupervised and combined model is small, as the unsupervised model mostly predicts a higher uncertainty than the supervised model.

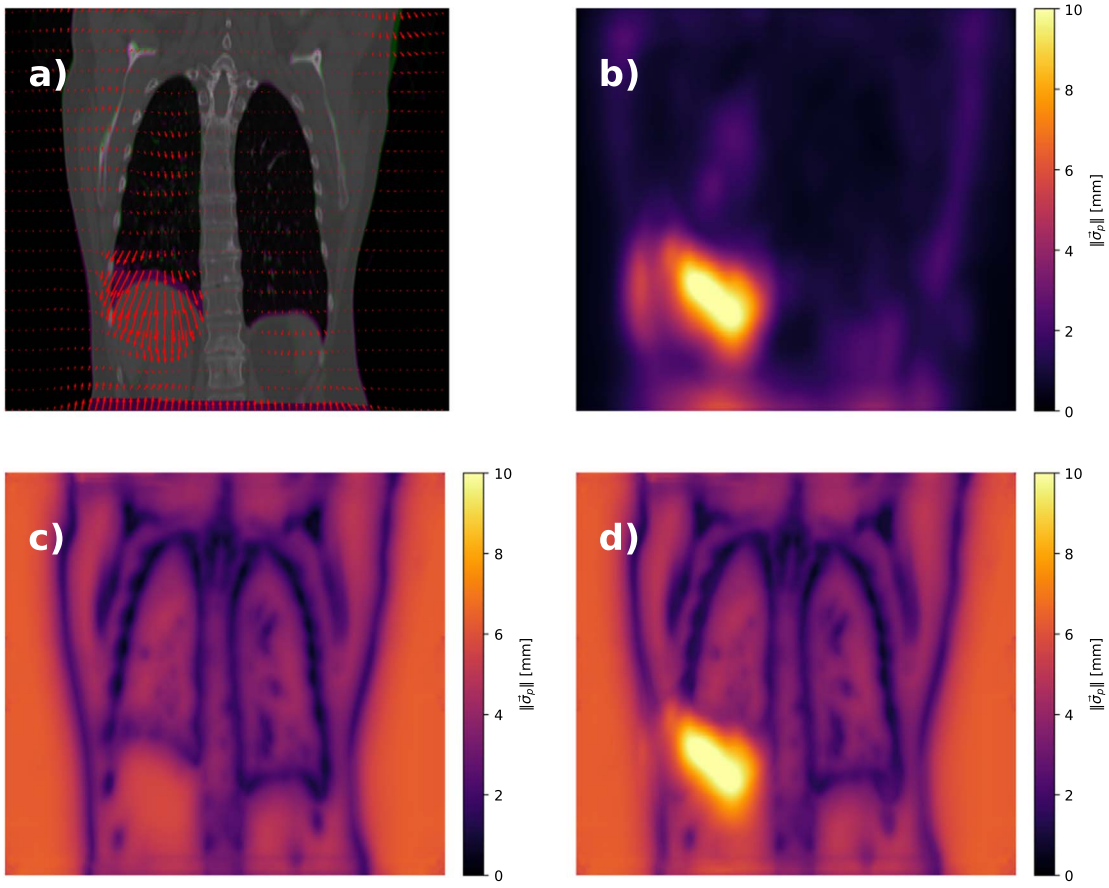

Figure 3. Visualization of the predicted uncertainties with the different models: (a) fixed (purple) and moving (green) image together with the mean DVF (red arrows); (b) supervised model; (c) unsupervised model; (d) combined model.

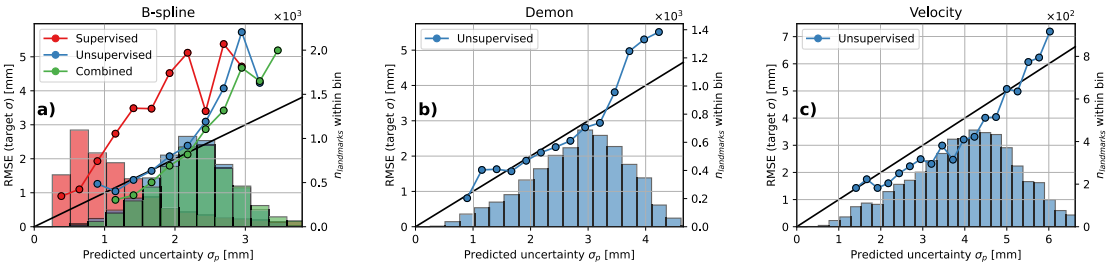

Figure 4. Landmark reliability diagram for the three models for b-spline (a) and the unsupervised model for demon and (b) velocity (c). The line plots represent the root-mean-squared error (RMSE, left axis), and the histograms depict how many landmarks were within the respective bin ($n_{\text{landmarks}}$, right axis).

Furthermore, the unsupervised model can predict with similar accuracy the uncertainty of the velocity DIR (LECE = 0.63 mm) and even with higher accuracy the uncertainty of demons (LECE = 0.41 mm) (figures 4(b) and (c)). For both DIRs, the hyperparameter tuning results in a larger uncertainty prediction than for b-spline, as the b-spline algorithm has a higher accuracy on the DIRLAB landmarks.

3.2.2. Probabilistic contour propagation

The propagated contour confidences are spread out over the interval [0, 1] where the contours are crossing regions with high DIR uncertainty (figure 5). For the NSCLC patients, probabilistic contour propagation with the supervised model is generally overconfident (figure 6(a)). For voxels with low predicted probabilities, i.e. voxels for which the model is relatively certain that they are outside the contour, a part was actually inside the contour. This part is larger than the predicted probability. The model is therefore overconfident, i.e. too certain that these voxels were outside the contour. Similarly, for voxels with high predicted probabilities, a part was actually outside the contour. This means that the uncertainty is underestimated, similar to what was found for

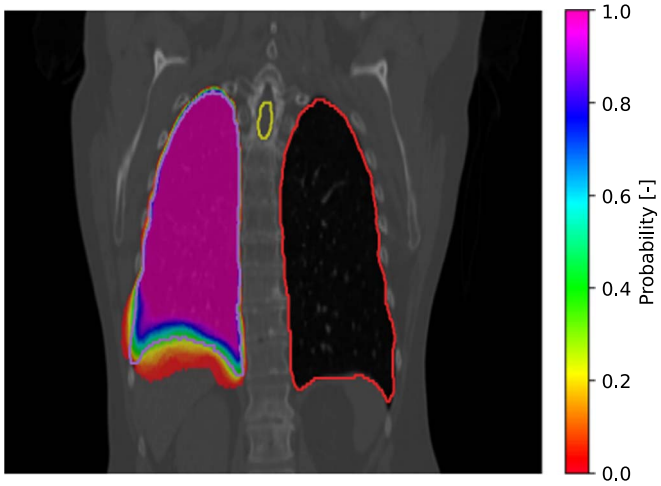

Figure 5. Visualization of the resulting propagated contour confidences of the left lung together with the manual contours for example in figure 3.

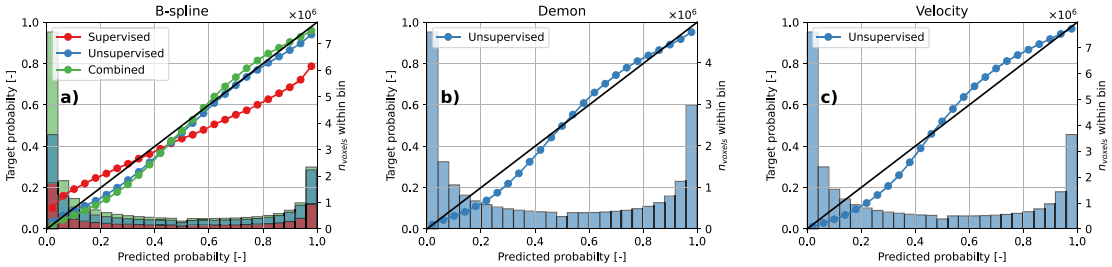

Figure 6. Contour reliability diagram for the three models for b-spline (a) and the unsupervised model for demon and (b) velocity (c). The line plots represent the actual proportion of voxels within the corresponding interval that were within the contour (target probability, left axis), and the histograms depict the number of voxels within the interval (n_{voxels}, right axis).

Table 2. Expected calibration error (ECE) for the different models [%]. The ECE is reported for the individual structures as well as averaged over all. The lungs were omitted as the manual contouring was inconsistent with regard to including the CTV into the lung structure or not.

		CTV	Esophagus	Heart	Spinal cord	**All**
B-spline	Supervised	15.5	9.3	9.0	10.4	**11.2**
	Unsupervised	7.7	1.6	2.9	5.5	**2.8**
	Combined	3.7	2.7	4.4	3.6	**2.6**
Demons	Unsupervised	7.1	3.8	4.3	5.7	**3.0**
Velocity	Unsupervised	5.6	3.1	6.0	6.0	**3.7**

the DIRLAB landmarks. The expected calibration error (ECE) is 11.2% on average, and even 15.5% for the CTV (table 2).

Contrarily, the unsupervised and combined models are slightly underconfident (figure 6(a)). For voxels with low probabilities (<0.5), the actual fraction of voxels inside the contour is even lower, i.e. model should have been more certain that these voxels were outside the contour. However, both models are much better calibrated than the supervised model, with average ECEs respectively 2.8% and 2.6% (table 2). That means that for a large enough group of voxels with confidence x, we can expect the proportion of voxels within the contour to be within $x \pm 2.8\%$. Even though the difference in average ECE between both models is small, the unsupervised model has a much larger ECE for the CTV (7.7% versus 3.7%) and spinal cord (5.5% versus 3.6%), showing the superiority of the combined model.

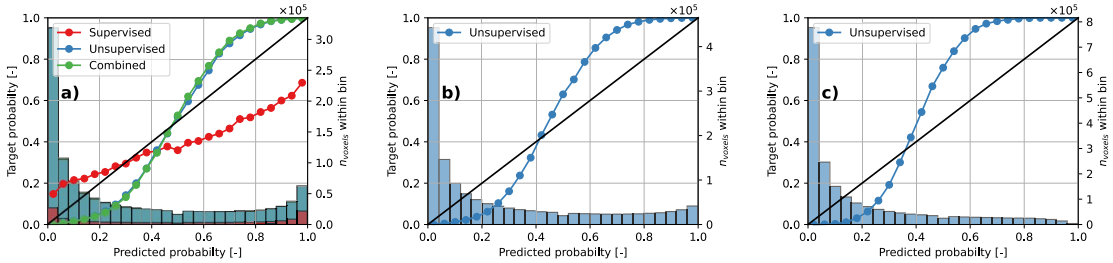

Figure 7. Contour reliability diagram for the brainstem for three models for b-spline (a), demons (b) and velocity (c). The line plots represent the actual proportion of voxels within the corresponding interval that were within the contour (target probability, left axis), and the histograms depict the number of voxels within the interval (n_{voxels}, right axis).

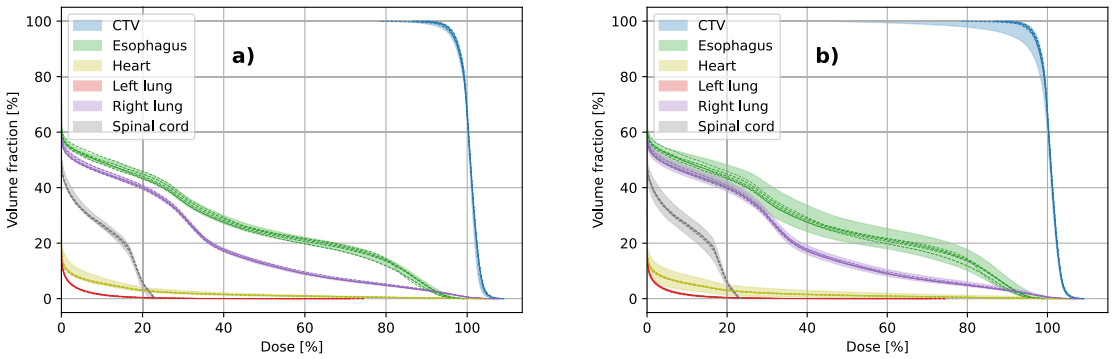

Figure 8. Example probabilistic DVHs (shaded areas) compared to the corresponding DVHs using 6 different DIR algorithms (dashed lines) for *not correlated* accumulation (a) and *fully correlated* accumulation (b).

Similarly, the tuned unsupervised models for demons and velocity are also underconfident but relatively well calibrated for contour propagation (figures 6(b) and c). Their ECEs are respectively 3.0% and 3.7%, slightly worse than the corresponding results for b-spline.

The total runtime of the algorithm, including model inference, generating 100 DVF samples and warping the 7 contours to the repeated CTs, was on average 37 s for the supervised and unsupervised models. For the combined model, inference has to be run twice, resulting in 44 s runtime.

Even though the models were tuned on deformation in the thorax region, their ECE averaged for all organs is only slightly worse for the HNC data (table C1 in the appendix). The supervised model again underestimates the uncertainty, but the unsupervised and combined models yield good results. However, the ECE for the individual structures varies significantly, indicating that the model is only well-tuned on average, but not for each individual organ. Specifically, the model overestimates the DIR uncertainty in regions with low contrast in the skull. The low contrast causes the unsupervised model to predict large DIR uncertainty, but since there is only little deformation, the registration is relatively well-defined. This effect is especially visible for the brainstem (figure 7). Therefore, before using the model on HNC data, the hyperparameters would have to be retuned.

3.2.3. Probabilistic dose accumulation

The predicted 95% confidence bounds of the probabilistically accumulated DVHs are largely dependent on the assumed correlation between the DVF samples of the repeated CTs (figure 8). Assuming that the samples are *not correlated*, dose differences cancel out, resulting in narrow bounds (figure 8(a)). Comparing these bounds with the accumulated DVHs for other DIR algorithms, we find that they are too narrow. For all models, the EVF of the accumulated DVH with other DIR algorithms is ranging between 29.9 and 62.4%, well below 95%, meaning that these bounds do not encompass such potential accumulated doses (table 3). Since the uncertainty prediction for a single DVF is well-tuned for both landmarks and contour propagation, this means that considering DVF samples as independent is inaccurate.

Contrarily, assuming that the samples are *fully correlated*, the DVF samples exhibit systematic errors which result in systematic dose differences, leading to wider DVH bounds (figure 8(b)). For all models and DIR algorithms, except the supervised model, the EVF is close to 95%, which is the target given the 95% confidence

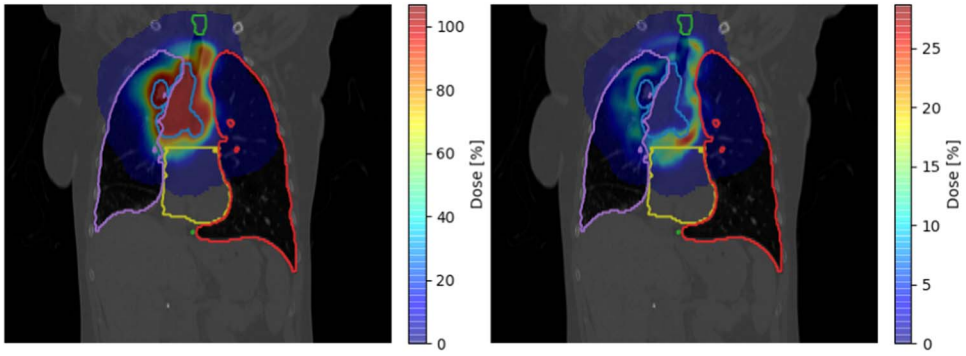

Figure 9. Mean accumulated dose (left) and standard deviation (right) for the patient depicted in figure 3.

Table 3. Average encompassed volume fraction (EVF) of each structure in the NSCLC dataset for which the accumulated DVH lies within the predicted accumulated *probabilistic* DVH [%]. The results are averaged over 5 DIR algorithms, excluding the 6th algorithm which was used to create the *probabilistic* DVH. The fractions are shown for each model and reference DIR algorithm, both for *not correlated* (NC) and *fully correlated* (FC).

			CTV	Eso-phagus	Heart	Left lung	Right lung	Spinal cord	**All**
B-spline	Supervised	NC	68.4	36.9	14.3	13.5	16.2	30.	**29.9**
		FC	86.3	66.2	24.	30.5	31.2	63.5	**50.3**
	Unsupervised	NC	61.1	77.2	63.5	36.9	48.9	75.8	**60.6**
		FC	93.8	97.6	99.6	90.8	94.2	95.5	**95.3**
	Combined	NC	56.9	80.2	62.8	42.8	48.2	83.4	**62.4**
		FC	96.7	98.	100.	97.3	98.7	96.4	**97.8**
Demons	Unsupervised	NC	32.2	66.2	58.7	27.6	45.6	65.5	**49.3**
		FC	82.8	98.3	100.	91.6	96.2	98.9	**94.6**
Velocity	Unsupervised	NC	29.2	59.	64.7	32.9	32.7	64.3	**47.1**
		FC	76.2	99.9	100.	97.2	99.9	98.9	**95.3**

bounds. The DVH bounds with the combined model are slightly too wide, as the EVF is larger than 95%. This means that the uncertainty is likely overestimated, similar to what was found for contour propagation.

Even though the *fully correlated* models on average capture the DVH uncertainty from accumulating with different DIR algorithms, the EVF differs for the individual structures. For example, the unsupervised model combined with velocity predicts too wide bounds for all OARs (EVF >95%) and too narrow ones for the CTV (EVF <95%), although yielding a well-tuned prediction on average. The effect is less pronounced for the other models and DIRs, but it shows the importance of evaluating the uncertainty on a structure-by-structure basis. Additionally, it can be visualized spatially, which shows that the accumulated dose uncertainty is high in regions where both the dose gradient and DIR uncertainty are high (figure 9).

The *probabilistic dose accumulation* for a single patient with 9 repeated CTs takes on average 3.5 min for the supervised and unsupervised networks: 91 s for running the 9 times model inference, 68 s for warping and accumulating the doses and 53 s for creating the *probabilistic DVH*. The combined model takes on average 5 min, as model inference has to be run twice.

Also for the HNC data, we find that the supervised model underestimates the DIR uncertainty (EVF $<<$ 95 %) and that fully correlated samples yield better results than not correlating (table C2 in the appendix). Even though the fully correlated unsupervised and combined models average EVFs close to 95%, the EVF for the individual organs clearly shows that the models are not well-tuned. For the OARs, the EVF is always close to 100%, showing that error bounds are too wide, i.e. overestimate the DIR uncertainty. This is in line with the results for contour propagation.

For the CTV, the EVF is too low. Interestingly, this is also caused by an overestimation of the DIR uncertainty. Inside the CTV, the dose is relatively homogeneous, and it drops only outside the CTV. When the DIR uncertainty is high, warping the dose at the CTV edge with DVF samples will sometimes result in the dose sampled inside the CTV and sometimes outside the CTV. Since the dose is homogeneous towards the inside of the CTV, and drops towards the outside of the CTV, the dose will only drop and never increase (contrary to

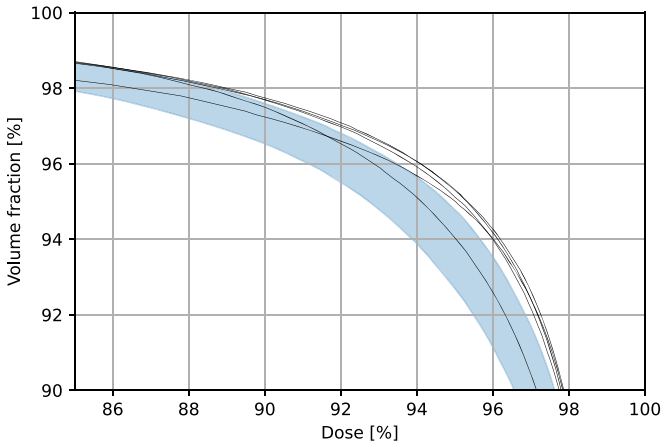

Figure 10. Zoom of the probabilistic DVH (shaded area) of the CTV of a HNC patient together with the DVHs of 5 different DIR algorithms (black lines).

OARs). Since the CTV is large, there will always be some regions on the edge where the dose warping yields a lower dose, causing the probabilistic DVH to shift downward in the low dose area (figure 10). It was verified that reducing the DIR uncertainty indeed increases the EVF, even though the width of the error bounds decreased. This again shows that the hyperparameters of the model should be retuned before using it for HNC.

4. Discussion

Despite the assumptions of Gaussian DVF uncertainty and a fixed correlation matrix, the unsupervised and combined models accurately predict landmark and contour uncertainty, and accumulated dose uncertainty resulting from running different DIR algorithms. Indeed, the predicted contour confidences deviate on average less than 3% from the expected values and the probabilistic DVHs encompass close to 95% of the DVHs accumulated with different DIR algorithms. This indicates that approximating the DIR uncertainty with a Gaussian is sufficient for practical applications, even though the underlying DIR uncertainty may not be Gaussian. Future work could investigate whether more complex distributions, such as skewed Gaussians, or learning the correlation matrix could further improve the performance.

The supervised model yields too low uncertainty estimates. This is because the uncertainty in the considered hyperparameters only accounts for part of the total DIR uncertainty, while other factors such as the transformation model and similarity criterion are ignored. This was further verified by generating G_{gt} for the DIRLAB scans, which showed that G_{gt} as defined in this work underestimates the landmark uncertainty (figure D1 in appendix D). The approach, however, is general. In future work, the MC sampling method used to generate G_{gt} can be expanded to include different DIR algorithms, transformation models, and similarity criteria, thereby accounting for a greater fraction of the DIR uncertainty.

The *expected calibration error* and *encompassed volume fraction* show that the unsupervised and combined models on average produce reliable uncertainty estimates. However, the metrics are worse when evaluating individual organs. This highlights a general problem for uncertainty evaluation: averaging results over many instances can conceal inaccuracies on the individual level. For example, in a hypothetical case with only two voxels, one in the middle of a contour and one very far from the contour, a model that predicts 50% confidence for both voxels would be considered perfectly calibrated (ECE = 0), even though these confidences are intuitively highly unlikely. Although our models would not make such predictions and many more voxels are considered, a similar averaging cannot be avoided during evaluation as it is inherently linked to quantifying uncertainty. It is therefore important to also assess these metrics for individual organs. In this case, this shows the superiority of the combined model over the unsupervised model, since the ECEs and EVFs for the individual organs are more uniform.

The advantage of the unsupervised model is that, even though it is trained with b-spline, it can be used to predict the uncertainty of other DIR algorithms. Indeed, for all applications studied here, the quality of the models for demons and Velocity is similar to the one for b-spline. This suggests that other DIR algorithms could use this network without the need for retraining, it would only require finding the optimal hyperparameters out of a set of already converged models. To tune the hyperparameters, we trained models with a wide variety of hyperparameters, so such converged models are already available and easy to share.

The unsupervised (and combined) models are slightly underconfident for contour propagation, meaning that the predicted uncertainties are a bit too large. This is because the model is tuned on the DIRLAB landmarks, and not on the NSCLC data itself. Since some DIRLAB scans contain larger deformations than the NSCLC scans, the DIR uncertainty is generally larger, so tuning on the DIRLAB scans results in a slight overestimation of the uncertainty for the NSCLC scans. This means that the models are on the conservative side, which is likely best for radiotherapy in general. However, by adjusting the hyperparameters, the predicted uncertainty can be decreased resulting in a lower ECE, which indicates that retuning the hyperparameters for a specific task or anatomy can improve the uncertainty estimation.

The runtime of the *probabilistic contour propagation* is below one minute and therefore suitable for online adaptive therapy. This means that future work can aim to use the predicted voxel-wise contour probabilities in combination with robust optimization, enabling the use of DIR for contouring in adaptive therapy. *Probabilistic dose accumulation* takes several minutes, which is significantly longer, as the inference and sampling procedure needs to be executed for each repeated scan. However, as part of an adaptive therapy workflow, this task can be performed offline and hence a runtime of several minutes is acceptable for clinical use in adaptive therapy. Also, the results can be stored after each fraction, requiring only the daily dose to be added, which would require approximately only two minutes each day.

The results clearly show that *fully correlating* DVF samples from different time points is more accurate than *not correlating* them. However, the accumulation only consisted of 9 daily doses, which is less than normally fractionated treatments. If there is a random component in the DVFs from different time points in most DIR algorithms, this random component will cancel out more when accumulating over more fractions. Therefore, future work should verify whether *fully correlating* DVF samples also yields realistic DVH bounds when accumulating over a more realistic number of fractions. Additionally, other models for temporal correlation should be tested, such as partially correlating a sample to the sample of the previous time point or linearly decreasing correlation.

The evaluation of the *probabilistic dose accumulation* relies on the assumption that the solutions of other DIR algorithms are potential samples of the deformation distribution, since the ground truth is unknown. The high EVF we find therefore mainly means that we can predict the uncertainty of such samples, rather than the underlying deformation uncertainty. The analysis could be extended in future work by the use of (digital) phantom data, where the ground truth deformation is known. Whereas we believe that this is an important extension, the results of such a study should be treated with care. Similar to our assumption that several DIRs span the actual deformation uncertainty, using phantom data requires the assumption that the phantom deformations span the actual patient deformations. Digital phantoms mostly rely on simplified deformation models, and therefore, might not cover all the modes of deformation plausible in a patient.

The accuracy of intensity-based DIR algorithms depends on the regularization strength. The optimal strength depends on the tissue type (Zhang *et al* 2021) or even the application (Kirby *et al* 2013). The performance of a model tuned for one anatomical region therefore generally drops when applied to another one. Similarly, our model was tuned for thorax deformation, and evaluation in the head and neck region shows that the performance drops. More specifically, our results indicate an overestimation of the predicted uncertainty. This is not surprising, as deformation in the thorax is generally larger which causes the DIR uncertainty to be larger. This means that the model is not directly applicable to other anatomical regions. In future work, this downside could be overcome by changing the hyperparameters of the unsupervised network. By either increasing λ or decreasing σ_I, the predicted uncertainty decreases which might improve the results on the HNC data. Since the unsupervised model has already been trained for a wide range of hyperparameters in this work, using it for another anatomy would not require retraining, only the evaluation of the ECE and EVF for the library of already converged models.

Finally, our work focuses on CT-to-CT registration. However, in most cases, a daily CBCT or MRI is acquired instead of a daily CT. Our approach could remain useable in case such daily images are converted into a synthetic CT, which is anyway mostly necessary for reoptimization of the treatment plan. However, as the quality of synthetic CTs differs from actual CTs, future work should validate whether the DIR uncertainty prediction also works for synthetic CTs. Additionally, the loss function could be adjusted for intramodality registration, which would allow e.g. probabilistic contour propagation from CT to CBCT, or even probabilistic generation of a synthetic CT. For the supervised model, this would simply require running b-spline intermodality with different hyperparameters. For the unsupervised model, this could be attempted by changing the first term in equation (7) by an intermodality similarity criterion, such as mutual information.

5. Conclusion

In this work, deep learning models were developed to quantify DIR uncertainty to estimate propagated contour and accumulated dose uncertainty. The results show that the unsupervised and combined models predict the DIR uncertainty more accurately than the supervised model in the thorax region. Evaluating the propagation of the uncertainty predicted by these models into contour uncertainty yields expected calibration errors of respectively 2.8% and 2.6%. Furthermore, propagation of DIR uncertainty into probabilistic accumulated DVHs yields encompassed volume fractions close to the expected 95%. Combined with acceptable runtimes, this demonstrates that the models are promising for use in adaptive radiotherapy, even though hyperparameter retuning and validation will be necessary when the model is used for different anatomical regions or in combination with other DIR algorithms.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Marie Skodowska-Curie Actions under Grant Agreement No. 955956. The authors would like to thank Enrique Amaya, Marc Walser, Barbara Bachtiary and Reinhardt Krcek for contouring of daily CTs in the NSCLC and HNC datasets. We would further like to thank Cosylab for providing their DIR algorithm.

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. The code and trained models regarding this work can be found at: https://github.com/AndreasSmolders/DIRUncertainty.

Appendix A. Model input

All models in this work predict the DVF uncertainty based on three inputs: the fixed image, the moving image and a (mean) DVF calculated by another DIR algorithm. This DVF is included in the input for two main reasons:

(i) If the DVF were not included, the networks would not know where the mean field is exactly pointing. That would also mean that it does not know how well the fixed and the moved image match given the mean field, and it could therefore not as accurately predict the uncertainty of the DIR.

(ii) The DVF contains the magnitude of the deformation. Especially for the supervised model, we find that the predicted uncertainty is very strongly correlated with this magnitude, and therefore the model could again not predict as accurately the DIR uncertainty.

This last point can be shown by comparing the results of the supervised network with and without the DVF as input (table A1).

Table A1. Mean absolute error (MAE) for the supervised model with and without the mean DVF as input. The results are shown for the test and validation set of the *CPT dataset*, containing each approx. 10% of the data.

	With DVF input	Without DVF input
Validation MAE [voxels]	0.14	0.18
Test MAE [voxels]	0.14	0.21

Appendix B. Hyperparameter tuning

The hyperparameters of the supervised model consist of the parameters of the probability distributions of the b-spline regularization λ_r and grid spacing s, and need to be tuned before training. B-spline was run for the validation scans of the DIRLAB dataset for a range of λ_r and s, and for each set the target registration error was evaluated (figure B1).

The TRE is minimal for $\lambda_r = 0.1$. It increases strongly for $\lambda_r > 1$, but is not strongly affected by $\lambda_r < 0.01$, because in that case the regularization does not affect the solution anymore. This leads to the choice of distribution $\log_{10} \lambda_r \sim \mathcal{N}(-1.5,\ 1.25)$ (figure B1(a)). The optimal grid spacing is $s = 13$ mm. The TRE increases strongly when s is decreased, and increases much slower when s is increased. This leads to the choice of distribution $s \sim \mathcal{U}(5,\ 50)$ mm (figure B1(b)). For both hyperparameters, the bounds were purposely set relatively wide, as these could be necessary for other anatomical regions or magnitudes of deformation.

Unsupervised training was repeated for several combinations of the hyperparameters λ and σ_I^2 and the posterior probability of the DIRLAB landmarks $p(LM)$ of the validation set was evaluated (figure B2(a)). For b-spline, it is maximal when $\lambda = 20$ and $\sigma_I^2 = 2 \cdot 10^{-4}$.

Since the model is unsupervised, it should principally be able to predict the uncertainty of any DVF, not only of those resulting from a b-spline algorithm, and DVFs from other algorithms can also be used as input. However, depending on the quality of that algorithm, different hyperparameters are optimal. Evaluating the $p(LM)$ for demons, we find that the optimal hyperparameters are $\lambda = 10$ and $\sigma_I^2 = 2 \cdot 10^{-4}$. For velocity, $\lambda = 20$, $\sigma_I^2 = 10^{-3}$ and $\lambda = 5$, $\sigma_I^2 = 3 \cdot 10^{-4}$ yield similar optimal results, and the latter is used in the following.

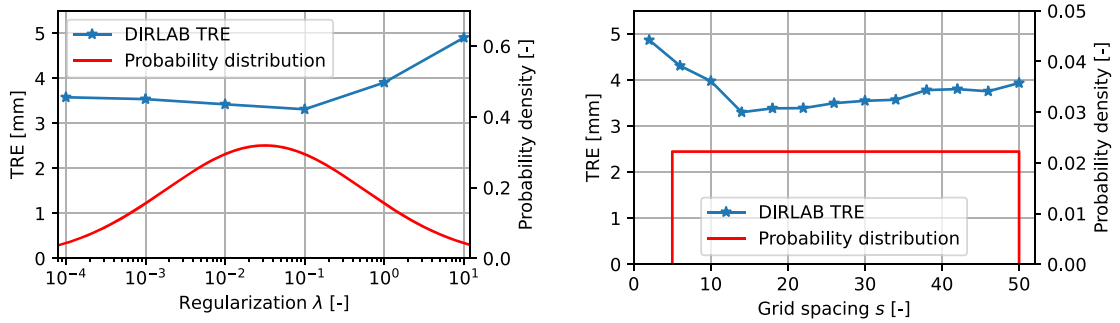

Figure B1. Target registration error TRE for the landmarks in the DIRLAB validation set for different values of the regularization λ_r (left) and grid spacing s (right). For the left plot, the TRE is averaged over all values of s, and for the right one averaged over λ_r. The corresponding selected probability densities are visualized on the right axis.

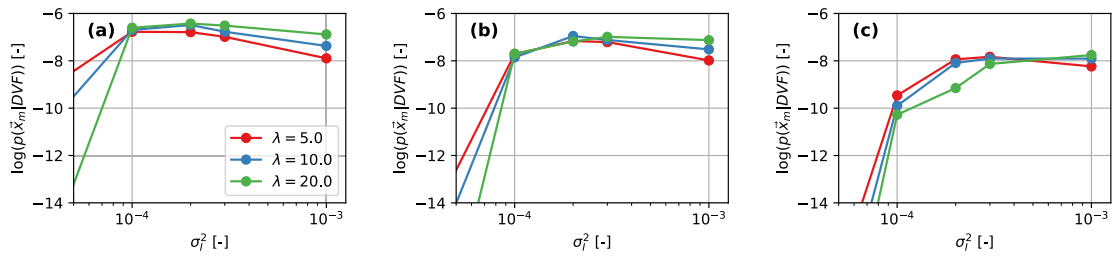

Figure B2. Mean log of the posterior probability of observing the moving DIRLAB landmarks for different values of the hyperparameters λ and σ_I^2 for unsupervised learning. Results are shown individually for (a) b-spline, (b) demon and (c) velocity.

Appendix C. Results head and neck cancer patients

C.1. Probabilistic contour propagation

Table C1. Expected calibration error (ECE) for the different models in the HNC dataset [%]. The ECE is reported for the individual structures as well as averaged overall.

		CTV	CTV boost	Spinal cord	Brainstem	Chiasm	Eyes	Parotids	Optic nerves	All
B-spline	Supervised	15.2	16.4	14.1	17.1	19.8	13.7	18.1	12.9	**15.4**
	Unsupervised	2.5	5.7	3.5	7.8	2.8	3.6	6.9	3.1	**2.4**
	Combined	2.5	4.6	2.0	8.2	2.7	3.7	5.6	3.3	**2.5**
Demons	Unsupervised	4.4	4.9	3.6	8.8	2.0	6.5	4.0	3.7	**4.1**
Velocity	Unsupervised	4.8	4.5	1.5	9.3	1.4	7.6	4.3	3.6	**4.7**

C.2. Probabilistic dose accumulation

Table C2. Average encompassed volume fraction (EVF) of each structure in the HNC dataset for which the accumulated DVH lies within the predicted accumulated *probabilistic* DVH [%]. The results are averaged over 5 DIR algorithms, excluding the 6th algorithm which was used to create the *probabilistic* DVH. The fractions are shown for each model and reference DIR algorithm, both for *not correlated* (NC) and *fully correlated* (FC).

			CTV	CTV boost	Spinal cord	Brainstem	Chiasm	Eyes	Parotids	Optic nerves	All
B-spline	Supervised	NC	64.7	95.9	51.1	24.3	43.0	37.5	32.2	43.0	**45.8**
		FC	79.7	97.0	74.6	44.3	61.4	66.7	56.9	63.0	**66.4**
	Unsupervised	NC	53.6	81.0	81.9	99.3	82.3	100.0	90.0	91.9	**87.4**
		FC	86.0	99.0	98.1	100.0	100.0	100.0	100.0	100.0	**98.5**
	Combined	NC	48.3	76.5	82.0	99.4	78.2	98.6	92.4	92.1	**86.4**
		FC	76.7	99.0	99.6	100.0	99.9	100.0	100.0	99.3	**97.6**
Demons	Unsupervised	NC	50.2	80.0	86.7	99.5	93.8	100.0	94.4	94.7	**89.9**
		FC	81.9	98.9	100.0	100.0	100.0	100.0	99.8	100.0	**98.2**
Velocity	Unsupervised	NC	31.4	53.8	92.2	99.4	98.9	97.8	89.5	90.3	**84.6**
		FC	58.7	98.3	99.9	100.0	100.0	100.0	100.0	99.9	**96.1**

Appendix D. G_{gt} for DIRLAB scans

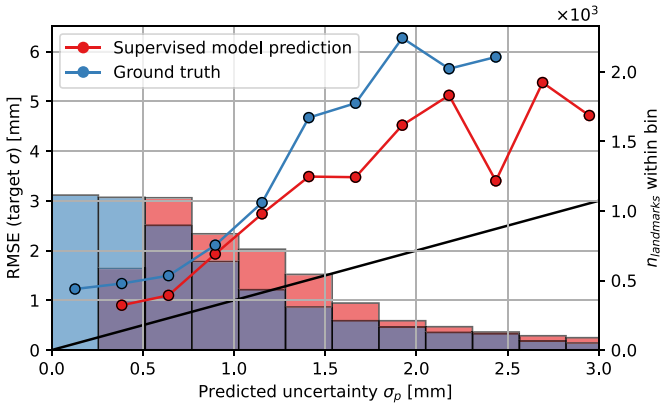

Figure D1. Comparison of the DIRLAB landmark uncertainty between the supervised model prediction and G_{gt} derived by creating 100 samples varying the b-spline parameters.

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

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

Balakrishnan G *et al* 2019 VoxelMorph: a learning framework for deformable medical image registration *IEEE Trans. Med. Imaging* **38** 1788–800

Bortfeld T 2006 IMRT: a review and preview *Phys. Med. Biol.* **51** R363–79

Brock K K *et al* 2005 Accuracy of finite element model-based multi-organ deformable image registration *Med. Phys.* **32** 1647–59

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

Castillo E *et al* 2010 Four-dimensional deformable image registration using trajectory modeling *Phys. Med. Biol.* **55** 305–27

Castillo R *et al* 2009 A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets *Phys. Med. Biol.* **54** 1849–70

Çiçek Ö *et al* 2016 3D U-Net: learning dense volumetric segmentation from sparse annotation *Medical Image Computing and Computer-assisted Intervention—MICCAI 2016* ed S Ourselin *et al* (Springer International Publishing) pp 424–32

Dalca A V *et al* 2019 Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces *Med. Image Anal.* **57** 226–36

de Vos B D *et al* 2019 A deep learning framework for unsupervised affine and deformable image registration *Med. Image Anal.* **52** 128–43

Guo C *et al* 2017 On calibration of modern neural networks *Proc. of the 34th Int. Conf. on Machine Learning* (*Proc. of Machine Learning Research, PMLR*) vol 40ed D Precup and Y W Teh pp 1321–30

Hardcastle N *et al* 2013 Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy *Radiat. Oncol.* **8** 243

Heinrich M P *et al* 2016 Deformable image registration by combining uncertainty estimates from supervoxel belief propagation *Med. Image Anal.* **27** 57–71

Hering A *et al* 2023 Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning *IEEE Trans. Med. Imaging* **42** 697–712

Jaderberg M *et al* 2015 Spatial transformer networks *Advances in Neural Information Processing Systems* ed C Cortes *et al* (Curran Associates, Inc.) vol 28

Kadoya N *et al* 2016 Multi-institutional validation study of commercially available deformable image registration software for thoracic images *Int. J. Radiat. Oncol.***Biol.***Phys.* **96** 422–31

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization arXiv:1412.6980

Kingma D P, Salimans T and Welling M 2015 Variational dropout and the local reparameterization trick *Advances in Neural Information Processing Systems* ed C Cortes *et al* (Curran Associates, Inc.) vol 28

Kirby N *et al* 2013 The need for application-based adaptation of deformable image registration *Med. Phys.* **40** 011702

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 Elastix: a toolbox for intensity-based medical image registration *IEEE Trans. Med. Imaging* **29** 196–205

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

Liu W *et al* 2012 Robust optimization of intensity modulated proton therapy *Med. Phys.* **39** 1079–91

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

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

Murr M *et al* 2023 Applicability and usage of dose mapping/accumulation in radiotherapy *Radiother. Oncol.* **182** 109527

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

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

Rigaud B *et al* 2019 Deformable image registration for radiation therapy: principle, methods, applications and evaluation *Acta Oncol.* **58** 1225–37

Rueckert D *et al* 1999 Nonrigid registration using free-form deformations: application to breast MR images *IEEE Trans. Med. Imaging* **18** 712–21

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)*

Simpson I J A *et al* 2013 Ensemble learning incorporating uncertain registration *IEEE Trans. Med. Imaging* **32** 748–56

Smolders A *et al* 2023a Dosimetric comparison of autocontouring techniques for online adaptive proton therapy *Phys. Med. Biol.* **68** 175006

Smolders A *et al* 2022a Deformable image registration uncertainty quantification using deep learning for dose accumulation in adaptive proton therapy *Biomedical Image Registration* ed A Hering (Springer) pp 57–66

Smolders A *et al* 2022b Fast deformable image registration uncertainty estimation for contour propagation in daily adaptive proton therapy *Medical Imaging with Deep Learning*

Smolders A *et al* 2023b Patient-specific neural networks for contour propagation in adaptive radiotherapy *Phys. Med. Biol.* **68** 095010

Smolders A *et al* 2023 Inter- and intrafractional 4D dose accumulation for evaluating ΔNTCP robustness in lung cancer *Radiother. Oncol.* **182** 109488

Sotiras A, Davatzikos C and Paragios N 2013 Deformable medical image registration: a survey *IEEE Trans. Med. Imaging* **32** 1153–90

Thirion J P 1998 Image matching as a diffusion process: an analogy with Maxwell's demons *Med. Image Anal.* **2** 243–60

Tran A et al 2017 Treatment planning comparison of IMPT, VMAT and 4π radiotherapy for prostate cases *Radiat. Oncol.* **12** 10
Unkelbach J et al 2018 Robust radiotherapy planning *Phys. Med. Biol.* **63** 22TR02
Weistrand O and Svensson S 2015 The ANACONDA algorithm for deformable image registration in radiotherapy *Med. Phys.* **42** 40–53
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
Zhang Y et al 2021 Tissue-specific deformable image registration using a spatial-contextual filter *Comput. Med. Imaging Graph.* **88** 101849