Physica Medica 112 (2023) 102642

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Exploring contrast generalisation in deep learning-based brain MRI-to-CT synthesis

Lotte Nijskens [a,b], Cornelis A.T. van den Berg [a,b], Joost J.C. Verhoeff [b], Matteo Maspero [a,b,*]

[a] Computational Imaging Group for MR Diagnostics & Therapy, Center for Image Science, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584CX, The Netherlands

[b] Department of Radiotherapy, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584CX, The Netherlands

ARTICLE INFO

Keywords:
Medical imaging
Radiotherapy
Artificial intelligence
Machine learning
Regression
Magnetic resonance imaging
Computed tomography
Generalisation
Domain shift

ABSTRACT

Background: Synthetic computed tomography (sCT) has been proposed and increasingly clinically adopted to enable magnetic resonance imaging (MRI)-based radiotherapy. Deep learning (DL) has recently demonstrated the ability to generate accurate sCT from fixed MRI acquisitions. However, MRI protocols may change over time or differ between centres resulting in low-quality sCT due to poor model generalisation.

Purpose: investigating domain randomisation (DR) to increase the generalisation of a DL model for brain sCT generation.

Methods: CT and corresponding T_1-weighted MRI with/without contrast, T_2-weighted, and FLAIR MRI from 95 patients undergoing RT were collected, considering FLAIR the unseen sequence where to investigate generalisation.

A "Baseline" generative adversarial network was trained with/without the FLAIR sequence to test how a model performs without DR. Image similarity and accuracy of sCT-based dose plans were assessed against CT to select the best-performing DR approach against the Baseline.

Results: The Baseline model had the poorest performance on FLAIR, with mean absolute error (MAE) = 106 ± 20.7 HU (mean ±σ). Performance on FLAIR significantly improved for the DR model with MAE = 99.0 ± 14.9 HU, but still inferior to the performance of the Baseline+FLAIR model (MAE = 72.6 ± 10.1 HU). Similarly, an improvement in γ-pass rate was obtained for DR vs Baseline.

Conclusion: DR improved image similarity and dose accuracy on the unseen sequence compared to training only on acquired MRI. DR makes the model more robust, reducing the need for re-training when applying a model on sequences unseen and unavailable for retraining.

1. Introduction

Radiation therapy (RT) plays a crucial role in cancer treatment, benefiting approximately half of all cancer patients [1]. Computed tomography (CT) is the primary imaging modality for RT planning as it provides accurate electron density information required for dose calculations [2]. Magnetic resonance imaging (MRI), on the other hand, offers superior soft tissue contrast compared to CT and has been proposed as the preferred modality for delineating tumors and surrounding organs at risk (OARs) [3]. MRI has proven valuable in reducing variability in tumor and OAR delineations across various disease sites, such as breast [4], prostate, and head-and-neck cancers [5]. Moreover, for certain brain cancer patients, MRI plays a crucial role in identifying tumor boundaries that may not be clearly visible on CT [6,7].

While CT provides the necessary information for dose calculations, MRI lacks the inherent electron density characteristics required in RT [8]. As a result, CT and MRI images are often acquired and fused for RT planning [9], potentially introducing uncertainties due to misregistrations [10,11]. MRI-only RT offers several advantages [12,13], including reduced patient exposure to ionising radiation, particularly beneficial in scenarios requiring re-planning [14] or for pediatric populations [15]. Additionally, MRI-only RT improves patient comfort by reducing the number of scans required and simplifies the workflow, leading to reduced workload [16–18] and costs [16,17,19]. With the introduction of MRI-guided RT [20,21], the interest in MRI-only RT has grown significantly [22,23]. Commercial products are available to

* Corresponding author at: Computational Imaging Group for MR Diagnostics & Therapy, Center for Image Science, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584CX, The Netherlands.
E-mail address: m.maspero@umcutrecht.nl (M. Maspero).
URL: https://compimag.org/members/matteo-maspero (M. Maspero).

https://doi.org/10.1016/j.ejmp.2023.102642
Received 18 March 2023; Received in revised form 24 May 2023; Accepted 5 July 2023
Available online 18 July 2023

facilitate MRI-only in the prostate, brain, and head and neck, making MRI-only a clinical reality [24].

The main obstacle in implementing MRI-only RT is the lack of a direct relationship between the nuclear magnetic properties of tissues and their electron density characteristics for dose calculations. Numerous approaches have been proposed to represent MRI as a CT equivalent to overcome this, resulting in synthetic CT (sCT) images [24–26].

Recently, deep learning (DL)-based methods became of particular interest as their inference requires limited time (seconds to minutes) for sCT generation, unlike classical image processing-based methods (ten minutes to hours) [27,28]. This time aspect is essential in MRI-guided RT, requiring sCT generation within a few minutes to allow daily re-planning [22,25].

However, DL models have limitations in generalising to new domains [29,30]. DL models assume a shared statistical distribution and feature space between the training and test data, necessitating retraining if the test data lies outside the distribution [31]. In the case of sCT generation, new domains can include different MRI sequences, images acquired in different hospitals with varying acquisition parameters, scans acquired after a scanner upgrade or with a different model, or even images of different anatomies.

Most existing DL models for sCT generation have been trained and tested on specific anatomical sites using a limited set of MRI sequences and fixed imaging parameters [24]. These models often overlook the variability in MRI acquisition protocols used in clinical practice or the potential protocol changes over time. Robust and general models capable of producing sCT images from previously unseen MRI sequences would greatly facilitate the widespread clinical implementation of MRI-only RT [32,33]. However, achieving this level of generalisation remains a challenge for deep learning techniques [34,35].

Recent studies have investigated the generalisation of DL models to multiple MRI sequences in the context of MRI-only RT [36,37]. These studies attempted to improve inference performance by retraining the network on the additional sequences. However, poor generalisation was observed when the models were evaluated on sequences not included in the training data [36,37].

Recently, a promising technique called domain randomisation was proposed to improve a segmentation network's ability to generalise to unseen MRI sequences [38,39]. The method relies on the hypothesis that increasing variability in synthetic training data forces the model to provide accurate output for all domains [40], e.g., in MRI sequences and is motivated by experiments showing that data augmentation beyond realism improves generalisation [41]. In this work, we explore domain randomisation to develop DL-based models for MRI-to-sCT generation that can generalise to MRI scans acquired with unseen sequences in brain MRI-only RT. Inspired by previous work by Billot et al. [38,39], we propose a domain randomisation method that generates synthetic images with random contrasts to enhance contrast generalisation. The underlying hypothesis is that training a DL model for MRI-to-sCT generation on input data with synthetic, rather than necessarily realistic, image contrasts compels the network to learn contrast-agnostic features [40,41]. To our knowledge, domain randomisation has not been applied yet to MRI-only RT.

We investigate two approaches to domain randomisation: (1) training on images with synthetically generated random contrast derived from segmented MRIs, and (2) training on random linear combinations of multiple MRI sequences. The effects of domain randomisation are compared to training solely on a mixture of acquired sequences. Our goal is to explore the extent to which a DL model can become contrast-agnostic and capable of generating sCT images that enable clinically acceptable dose calculations for RT planning.

By conducting this study, we aim to contribute to developing robust DL models for MRI-to-sCT generation that can generalise to unseen MRI sequences. Such models can potentially advance the implementation of MRI-only RT and improve treatment planning accuracy in clinical practice.

2. Materials and methods

2.1. Data collection and imaging protocols

Data from 95 patients were selected undergoing treatment at the UMC Utrecht RT department from a large retrospective, anonymised database collected under the local Medical Ethical Committee's approval (study number: 20/519, approved on August 11, 2020). The main selection criterion was the availability of a treatment plan for brain RT conducted between January 2020 and July 2021, with corresponding CT and MRI (T_1-weighted with and without contrast enhancement, T_2-weighted and FLAIR images). Patients were excluded if not all sequences were available, no suitable CT was available, the time between MRI and CT acquisition exceeded 1.5 months, the patient's age was <18 years, or the MRI was a follow-up exam. If multiple CT acquisitions were available, the most recent one was chosen, and the MRI dataset acquired closest in time to the CT was selected.

Patients were randomly divided over the training (n = 60), validation (n = 10) and test set (n = 25). The female/male ratio for the 95 included patients was 51/44 with a mean patient age of 59.9 ±13.0 years (range: 24.3–86.8). In total, 66 patients were acquired on 1.5 T, and 29 on 3.0 T. The mean interval between CT and MRI acquisition was six days (range: min–max = 1–26). Dose prescriptions ranged from 14 to 60 Gy over 1–33 fractions of 2.0–3.0 Gy.

Planning CTs were acquired at the radiotherapy department using a Brilliance Big Bore system (Philips Healthcare, USA). The acquisition occurred in the supine treatment position, aided by head support and a personalised immobilisation mask. CT acquisition was without contrast agents, with a tube potential of 120 kV, a tube current of 234–360 mA (range = min–max), and 1000–1712 ms exposure. The in-plane resolution was 0.57–1.17 mm^2, with a slice thickness of 1–2 mm.

MRI data were acquired with a 1.5 or 3.0 T Ingenia MR-RT system (Philips Healthcare, the Netherlands). Available sequences (Table 1) were: 3D T_1-weighted turbo field echo (TFE) images with and without Gadolinium contrast (T1w and T1wGd), 2D T_2-weighted turbo spin-echo (TSE) images with Gadolinium contrast (T2w) and 3D T_2-weighted FLAIR TSE images (FLAIR).

2.2. Image processing

If not otherwise specified, image processing and performance evaluation was performed in Matlab R2019a (The MathWorks, Inc., USA).

Pre-processing Each MRI was rigidly registered to the corresponding CT with `Elastix` (version 4.700) [42,43], using multi-resolution registration (with a resolution of 4, 2, 1, and 0.5 times the reconstructed voxel size) with an adaptive stochastic gradient descent optimiser and mutual information similarity metric. The parameters from [44] were adopted. The registered MRI will be referred to as MRI_{reg}. MRI_{reg} and CT were resampled to isotropic 1.0 × 1.0 × 1.0 mm^3 resolution using linear interpolation.

Most images contained a discrepancy between CT and MRI FOVs, caused by angled MRI acquisition (Fig. 1). A body mask was computed on the non-registered MRI to ensure congruent FOVs between CT and MRI_{reg}. The mask was generated using a threshold with an empirically determined value of 20 (or 15 for T2w images), followed by morphological filling and dilation with a disk-shaped structuring element of radius 20 voxels. The binary mask was registered to the CT by applying the transform computed for MRI_{reg}, resampled, and applied to the MRI_{reg}-CT pair for training. The MRI_{reg} and CT FOVs were cropped to the extent of the registered mask with additional ten voxel margins on each side or until the original image boundary. MRI_{reg} were normalised by clipping to the per-patient 99th percentile over the masked volume. Training CTs were clipped to range [−1024, 1500] HU, but the range was kept untouched for evaluation purposes.

Table 1
Overview of acquisition parameters per sequence for the 95 included patients.

Parameter	3D T1w TFE	3D T1w TFE Gd	2D T2w TSE	3D T2w FLAIR TSE
B_0 [T]	**1.5** (66)	**1.5** (66)	**1.5** (66)	**1.5** (66)
	3.0 (29)	**3.0** (29)	**3.0** (29)	**3.0** (29)
Contrast	No	Yes	Yes	No
Read-out	AP	AP	AP	AP
Flip angle [°]	8	8	90	90
TR [ms]	7.6–8.7	7.6–8.7	3119–5996	4800
TE [ms]	3.5–4.1	3.5–4.1	80–100	303–363
FOV[a] [mm^3]	230, 160	230, 160	230, 140–160	230, 160
Acq voxel[a] [mm^3]	1.0, 0.5–1.0	1.0, 0.5–1.0	0.6–0.7, 4.0–5.0	1.1–1.2, 0.6
Rec voxel[a] [mm^3]	0.4–1.0, 0.5–1.0	0.5–1.0, 0.5–1.0	0.4–0.5, 4.0–5.0	1.0, 0.6
Rec matrix[a]	240–512, 162–323	240–480, 162–323	480–512, 31–43	240, 269–270
BW [Hz/px]	190–217	190–217	143–206	851–1075
Acq time [s]	136–271	121–271	117–137	331–475

Acq time: acquisition time; Acq voxel/Rec voxel: acquisition/reconstruction voxel size; AP: anterior-posterior; B_0: main magnetic field strength; BW: bandwidth; FOV: field-of-view; Rec matrix: reconstruction matrix; TE: echo time; TFE: turbo field echo; TR: repetition time; TSE: turbo spin echo.
[a]Directions: anterior-posterior, right-left and craniocaudal.

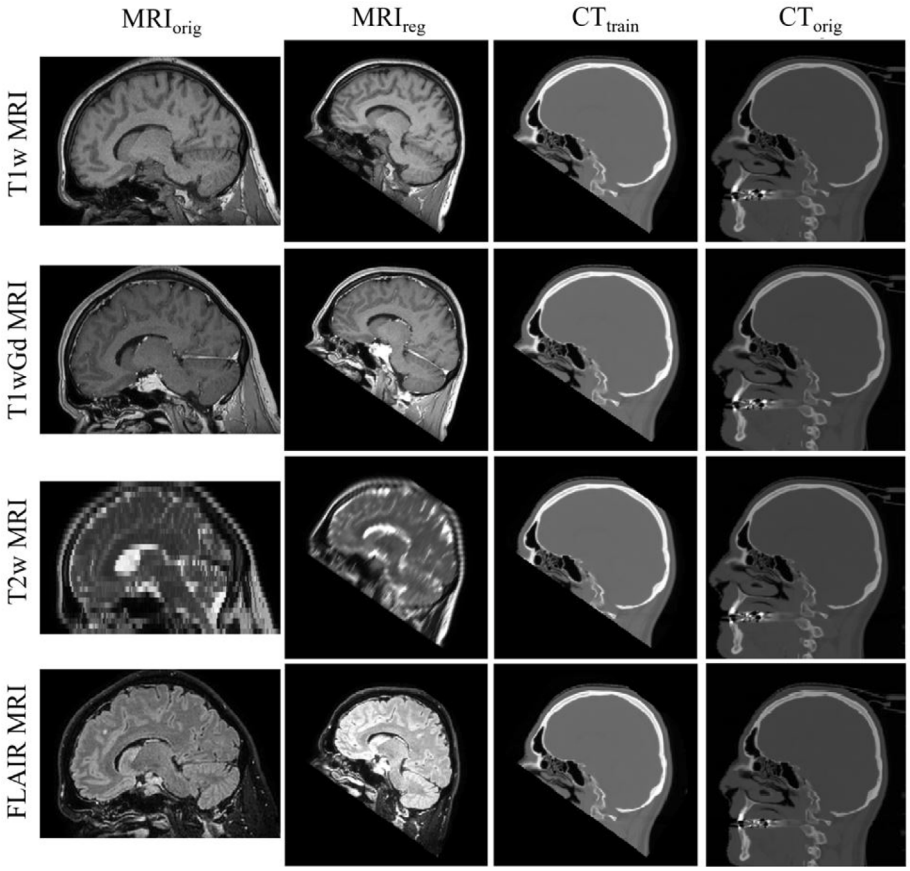

Fig. 1. Example of the pre-processing outcomes. The original T1w (top row), T1wGd (second row), T2w (third row), and FLAIR (bottom row) brain MRI for a single male patient in the training dataset are shown (left) with corresponding normalised MRI_{reg}, CT_{train} and ground truth CT_{crop} (left to right).

For CT, the masking and range clipping steps were only applied to the training images (hereafter: CT_{train}). CT_{train} and normalised MRI_{reg} were saved as 3D volumes in NifTI format, linearly rescaled to [−1, 1]. Fig. 1 shows an example of a normalised brain MRI_{reg} with the corresponding normalised CT_{train}, ground truth CT_{crop} and original, unregistered MRI for each sequence for a single patient.

Post-processing For post-processing, all generated sCTs were linearly rescaled to a [−1024, 1500] HU range, conforming to the range of CT_{train}.

2.3. Performance evaluation

The quality of the generated sCTs was evaluated in terms of image similarity between acquired CT and generated sCT and dose accuracy. Statistical comparisons were performed with Wilcoxon signed-rank tests, with p-values <0.05 regarded as statistically significant. Moreover, training and inference times are reported.

2.3.1. Image similarity

The accuracy of the assigned HU values was analysed with a voxel-wise comparison between CT_{crop} (ground truth) and sCT. A body contour mask was applied to CT_{crop} and sCT before calculating the metrics

and comparing on the intersection of the two masks. The masks were created by thresholding the (s)CT above −200 HU, then morphologically closing and filling the combined mask to include the nasal cavities. The mean absolute error (MAE) was computed per patient. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) were additional metrics. The range and mean ± standard deviation ($\mu \pm \sigma$) over all patients in the validation or test set were calculated for each metric.

2.3.2. Dose accuracy

The clinically optimised dose plan was re-calculated for the final models on (s)CT. Generated sCTs were registered and resampled to the original, non-cropped CT, allowing only translations. Multi-resolution registration was performed (with a factor 4, 2, 1, and 0.5 to the voxel size) with an adaptive stochastic gradient descent optimiser and mutual information. In case the registration quality was considered poor by an observer or failed, three resolutions were used instead of four (resolution 1: 4, 4, 2; resolution 2: 2, 2, 1; resolution 3: 1, 1, 0.5).

A segmentation of the body contour of the non-cropped CT was taken from the clinical treatment plan, and the voxels outside the original MRI FOV and inside this body contour were set to 0 HU. The difference in FOV between sCT and acquired CT was thus water-filled in both images. The water-filled sCT and acquired CT are referred to as sCT_{wf} and CT_{wf}.

Plans were volumetric modulated arc therapy (VMAT) photon plans with a single arc with a beam energy of 6.0 MV. They were calculated with a Monte Carlo algorithm on a 3 mm^3 grid with 3% uncertainty. Plan re-calculation was performed on (s)CT_{wf} using GPUMCD [45].

Dose accuracy was assessed through the calculation of the dose difference (DD) relative to the prescribed dose (D_{presc}) in the high-dose region (D >90% of D_{presc}) [24]:

$$DD = 100 * \frac{D_{CT} - D_{sCT}}{D_{presc}}\%, \quad (1)$$

with D the dose (in Gy) in the (s)CT_{wf}-based dose plan. Korsholm et al. [46] proposed a criterion for the clinical acceptability of DD: the DD compared to a CT-based dose plan should be <2% for 95% of the patients. In this work, a more conservative criterion was adopted. Individual sCTs were considered acceptable if the DD was <1%.

Dose-volume histograms (DVH) were analysed for differences in D_{median} and D_{max} between sCT- and CT-based plans for the following OARs: brainstem, optic chiasm, lenses, cochleae, and pituitary gland. Additionally, a 3D-γ global analysis was conducted [47]. For the computation of γ-pass rates, a 10% dose threshold was used, with 3%,3 mm, 2%,2 mm, and 1%,1 mm criteria. Heilemann et al. [48] demonstrated the ability to detect clinically unacceptable VMAT plans using a 90% γ-pass rate threshold for the 2%,2 mm criterion. Nevertheless, the absence of clinically significant dose differences was not guaranteed [48]. Considering that evaluation of γ-pass rates is adopted for quality assurance of delivered plans, where uncertainty is higher, we adopted stricter thresholds in this work: 95% and 99% for the 2%,2 mm, and 3%,3 mm criteria. The primary metric was the 95% dose threshold on γ 2%,2 mm.

2.4. Network architecture

The cGAN model `pix2pix` was implemented to allow paired training, as proposed by Isola et al. [49]. Initial investigations showed that 3D models outperformed 2D ones [50]. Therefore, only 3D models are reported in this work.

An implementation of the original `pix2pix` model [49] called `Ganslate` [51] in PyTorch version 1.10 was used for 3D models. According to the server availability, all models were trained on a GPU Tesla P100 PCIe 16 GB or V100 PCIe 32 GB (NVIDIA Corp., USA).

A 3D U-Net generator architecture that allows variable patch sizes as input was adopted, along with a 70 × 70 PatchGAN discriminator [49]. The $L1$-based loss function proposed in [49] was implemented.

2.5. Model optimisation

Hyperparameter optimisation was performed with a grid search strategy on a subset of the training set consisting of ten patients from the training set with only T1w images without contrast. The hyperparameters leading to the lowest average MAE in the validation set were adopted. SSIM and PSNR were calculated as additional metrics. The hyperparameter grid search space is detailed in Supplementary Material I.A.

One patient was retrospectively excluded from the validation set after observing the T2w MRI and CT registration failure. Validation of all models except those trained in the hyperparameter tuning stage was thus done on a nine-patient validation set.

As a final optimisation step, the ratio between T1w images with/without contrast and T2w images in the training set was balanced, and the batch size was fine-tuned (Supplementary Material I.B.). This step involved training a subset of fifteen patients from the training set. After balancing, the final training dataset (n = 60) contained 60 T2w, 30 T1w, and 30 T1wGd images.

After optimisation, all models were trained with Xavier initialisation, Adam optimiser, patch size = 128 × 128 × 128 voxels, batch size = 1, λ = 5000, number of downsampling steps = 5, and a constant learning rate of 0.001. A sliding window was used for patch combination with a patch overlap of 0.5, as also found in [27] and Gaussian blend mode. The Adam optimiser [52] was implemented with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ as momentum parameters and no weight decay.

Early stopping was applied to avoid overfitting, using the MAE as a decision criterion. The MAE was calculated for the sequences included for training in the body contours for the patients in the validation set. A combined MAE was computed as the average of the seen sequences. The first iteration for which this combined MAE did not improve for the following three iterations was selected, evaluating every 50,000 iterations. Supplementary Material I.C. illustrates the early stopping method.

2.6. Domain randomisation

Domain randomisation was applied by training on generating random contrast (RC) images starting from label maps or generating linear combinations (LC) of the acquired MRI. Both approaches have been investigated and are described in the following.

2.6.1. Random contrast

The domain randomisation strategy comprising RC images (Section 2.6) requires segmenting patients' MRIs. Automatic segmentations were generated from the T1w images, complemented by some structures labelled using CT_{train}.

Segmentation of cerebral structures was performed on T1w images using an open-source DL network called FastSurfer [53]. OARs were added by segmentation of T1w MRI with a previously in-house developed DL algorithm that is clinically adopted (unpublished and developed for clinical use as in [54]), based on the DeepMedic model [55]. The GTV was obtained from the clinical segmentation. Cerebrospinal fluid (CSF) was labelled using a combination of FastSurfer labels and clinical segmentations. Also, CT_{train} was segmented by thresholding to obtain labels for bone and soft tissue. The resulting label maps were stored as an additional dataset. More details on image segmentation, a lookup table with included labels, and an example of created label maps are reported in Supplementary Material II.

Label maps were converted to RC images for network training, as proposed by Billot et al. [39] using TorchIO library [56]. Specifically, after randomly selecting a segmentation from the training data, each label was assigned a Gaussian function with the mean and standard deviation chosen randomly from a uniform distribution with ranges of [10, 240] and [1, 25], respectively. These ranges were based on the

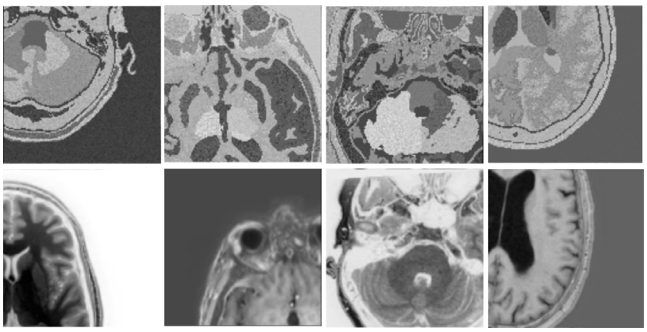

Fig. 2. Examples of random contrast (RC; top row) and linear combination (LC; bottom row) images generated from label maps. Each image is a slice from an example patch as input to the network during training.

sensitivity analysis in [39]. All voxels within a label were assigned an intensity value sampled from this Gaussian distribution.

Then, images were blurred to increase spatial coherence between neighbouring voxels. The standard deviation of the Gaussian was randomly sampled from a uniform distribution: $\sigma_{blur} \sim U(0, 0.3)$, like in [38]. Random gamma augmentation was applied after rescaling image intensity to a positive range to increase variability in the training data further. Following [39], the exponent $\gamma = e^{\beta}$ was randomly sampled from a normal distribution: $\beta \sim N(\mu_\beta, \sigma_\beta)$ with $\mu_\beta = 0$ and $\sigma_\beta = 0.4$. The RC image was rescaled to [−1, 1]; see Fig. 2(top) for an example of RC image patches.

2.6.2. Linear combination

Linear combination was considered to investigate a more straightforward method to perform domain randomisation than RC that would not require brain segmentations, facilitating the application of domain randomisation.

Linear combination (LC) images were generated from T1w(Gd) and T2w MRI. To enable linear combination, an additional training dataset was created in which the pre-processed T2w and T1wGd images were registered to the corresponding T1w image with the same registration parameters as in Section 2.2 A random choice was made during network training between combining the patient's T2w image with their T1w or T1wGd image, using equal probabilities. The T1w(Gd) and T2w images were then combined as follows:

$$Im_{LC} = p_1 * Im_{T1} + p_2 * Im_{T2}, \tag{2}$$

with p_1 and p_2 as the coefficients for voxel-wise addition of the T1w(Gd) (Im_{T1}) and T2w (Im_{T2}) image, respectively. These were randomly sampled from a uniform distribution: $p_{1,2} \sim U(-1, 1)$. The chosen range allows addition and subtraction in the linear combination and contrast inversions. Finally, images were rescaled to the range [−1, 1]. Fig. 2 (bottom) shows several examples of LC patches.

2.7. Experiment: random contrast vs linear combinations

An experiment was conducted to identify the most effective domain randomisation technique between RC and LC, comparing the two models in terms of image similarity on the validation set. Each model was trained with both the two domain randomisation approaches.

A model trained on a mix of acquired MRI and RC images derived from segmentations was adopted for this experiment to represent the RC method, after initial investigations showed that it outperformed a model trained on RC images only [50]. For model training, the entire training dataset was used. Hence, this RC model was trained on 60 segmentations, 30 T1w, 30 T1wGd, and 60 T2w images from the training set (n = 60 individual patients).

For the domain randomisation method comprising LC images, initial investigations indicated that a model trained with a 50% chance of applying a linear combination to the acquired MRI outperformed a model trained with a 100% chance of using a linear combination [50]. The LC model was thus trained with a 50% chance of linear combination. The dataset from which LC images were generated consisted of acquired T1w (n = 60) images and T1wGd (n = 60) and T2w (n = 60) images that had been registered to their T1w counterpart, as mentioned in Section 2.6.2 For the LC model, this LC-specific dataset and the original training dataset of 30 T1w images, 30 T1wGd images and 60 T2w images (Section 2.5) were used. A random choice was made at each iteration whether to apply a linear combination. The original dataset was sampled if an LC image should not be used.

The RC and the LC model were trained with the hyperparameters described in Section 2.5. Early stopping was based on the MAE obtained for sCT generated from T1w(Gd) and T2w images in the validation set for both models.

The RC and LC models were statistically compared using image similarity metrics calculated per sequence (T1w(Gd), T2w and FLAIR) on the validation set using MAE as the leading metric for model choice. The best-performing model was adopted as the final Domain Randomisation model.

2.8. Domain randomisation on an unseen sequence

In a final comparison, the chosen Domain Randomisation model that will result from the experiment described in 2.7 was compared to two models trained without domain randomisation: a Baseline model and a Baseline+FLAIR model. In this way, we can compare the impact of adding the unseen FLAIR sequence against domain randomisation compared to the baselines.

2.8.1. Baseline model

The Baseline model was trained on a mix of T1w, T1wGd, and T2w images to assess the models' ability to generalise to an unseen sequence (FLAIR) without domain randomisation. For model training, the entire training dataset was used: 30 T1w, 30 T1wGd, and 60 T2w images, applying the hyperparameters described in Section 2.5. The early stopping iteration was determined based on the MAE on T1w, T1wGd, and T2w images of patients in the validation set.

After determining when to apply early stopping on the validation set, the model was inferred on the test set (n = 25). Image similarity metrics and dose accuracy were calculated for sCT generated from patients' T1w, T1wGd, T2w, and FLAIR images in the test set. The image similarity metrics and metrics for dose evaluation were statistically compared between the four sequences. To discuss the dose accuracy on an individual patient level, we limit ourselves to the results obtained for the unseen FLAIR sequence.

2.8.2. Baseline+FLAIR vs Baseline

The Baseline+FLAIR model was trained to obtain a measure for the best achievable performance for FLAIR input images. This model was trained on the whole training set of 60 patients used for training the Baseline model, with the addition of FLAIR images. Hence, altogether the training dataset consisted of a mix of T1w (n = 30), T1wGd (n = 30), T2w images (n = 60), and FLAIR images (n = 60) from 60 individual patients. The hyperparameters described in Section 2.5 were adopted. For the Baseline+FLAIR model, the iteration for early stopping was determined based on the MAE on T1w, T1wGd, T2w, and FLAIR images of patients in the validation set.

After early stopping, the Baseline+FLAIR model was inferred on the test set (n = 25). Image similarity metrics and dose accuracy were calculated for sCT generated from each of the four sequences (T1w, T1wGd, T2w, and FLAIR). For each sequence, the image similarity metrics and metrics for dose evaluation were statistically compared with those obtained for the Baseline model. Also, statistical comparisons were made between the sequences.

Table 2
MAE obtained for sCT generated by the RC and LC models per MRI sequence. Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported. Wilcoxon-signed rank tests were used for statistical comparisons. Values of $p < 0.05$ were regarded as statistically significant.

Metric	Model	Sequence			
		T1w	T1wGd	T2w	FLAIR
MAE [HU]	RC	71.5 ± 12.1 [59.7 – 100]	69.6 ± 12.2 [56.6 – 98.6]	76.3 ± 10.9 [60.2 – 95.6]	105 ± 20.5 [74.1 – 142]
	LC	72.3 ± 12.4 [57.3 – 100]	71.0 ± 12.2 [58.2 – 99.8]	77.8 ± 11.4 [63.0 – 100]	110 ± 23.9 [72.9 – 155]
p-value		0.3	0.2	0.2	0.04

2.8.3. Domain randomisation vs baselines

For the final comparison, the Domain Randomisation model was inferred on the test set (n = 25 patients). As for the Baseline and Baseline+FLAIR model, image similarity metrics and dose accuracy were calculated for sCT generated from patients' T1w, T1wGd, T2w and FLAIR images. Per sequence, the image similarity and dose evaluation metrics were statistically compared to those obtained for the Baseline and Baseline+FLAIR model. Additionally, statistical comparisons were made between the sequences.

3. Results

For all the models inference time on the test set was approximately 4 s per sequence and patient.

3.1. Experiment: random contrast vs linear combinations

Early stopping was applied after 450,000 iterations for the RC model and 200,000 iterations for the LC model. For all sequences, the MAE obtained on the validation set was lower for the RC model than for the LC model (Table 2). Only the difference in FLAIR images was statistically significant: an MAE of 105 ± 20.5 HU was obtained for the RC model, compared to an MAE of 110 ± 23.9 HU for the LC model. Differences in SSIM and PSNR were not statistically significant (Supplementary Material III.A.).

Overall, using RC images was deemed the most beneficial domain randomisation strategy. Consequently, the RC model was adopted as the Domain Randomisation model for final comparison with the Baseline and Baseline+FLAIR models.

3.2. Domain randomisation on an unseen sequence

3.2.1. Baseline model

The training time for the Baseline model was 32.0 h, applying early stopping at iteration 300,000.

Among the four sequences, the performance of the Baseline model on the test set (n = 25) was best on T1w and T1wGd images (Fig. 3), with the difference between these two sequences not statistically significant for the three metrics. The p-values resulting from statistical tests of performance metrics (image similarity and dosimetric accuracy) between sequences are presented in Supplementary Material III.B. for each of the models. The mean MAE was 64.2 ± 7.3 HU or 63.8 ± 9.1 HU for T1w and T1wGd images, respectively. The worst performance was found for FLAIR images, with a considerable difference with performance on T1w(Gd) images: the mean MAE was 106 ± 20.7 HU. Testing on T2w images resulted in a mean MAE of 69.6 ± 8.5 HU. The difference in performance on FLAIR and T2w images compared to the performance on the other three sequences was statistically significant. Violin plots for SSIM and PSNR are shown in Supplementary Material III.C. Results for SSIM and PSNR were in line with the MAE for the Baseline model.

For the Baseline model, 3D γ-pass rates in the low dose region (>10% of the prescribed dose) with 1%,1 mm criterion were >95% (Table 3) for every patient and each sequence. The pass rate obtained for FLAIR images (99.0 ± 1.1%) was significantly lower than that computed for all other sequences. The γ-pass rates with 3%,3 mm and 2%,2 mm criteria were >99% for every patient and sequence. The obtained γ-pass rates with 3%,3 mm and 2%,2 mm criteria are shown in Supplementary Material III.D. for each of the models.

For the Baseline model, a DD in the high-dose region (>90% of the prescription dose) of −0.1 ± 0.2% was obtained for treatment plans based on sCT generated from T1w, T1wGd and T2w images, and a DD of 0.4 ± 0.5% was found for FLAIR images (Table 3). The DD was significantly larger for FLAIR than the three other sequences. Other differences in DD between sequences were not statistically significant.

The DD in treatment plans from FLAIR-based sCT was ≤1.5% and >1% for three patients (PT2, PT13, and PT18). Specifically, a discrepancy between sCT and CT HU values was found for PT2 near the high-dose region: the skull near the frontal lobe was too thinly on sCT, causing HU values to be lower than in the CT. Discontinuities were visible in the skull of this post-surgical patient in the problematic area, although no part of the skull had been resected. For PT13, the high-dose region was located in the dorsal part of the cerebrum, where differences in skull thickness occurred between the FLAIR-based sCT generated by the Baseline model and the acquired CT, this time with higher HU values in the sCT than in the acquired CT. Notable dose differences were observed for PT18 near the nasal cavities, close to one of the isocentres of irradiation. The sCT generated by the Baseline model from this patient's FLAIR image revealed more prominent differences between HU values of sCT and acquired CT than the sCT generated for the other sequences.

Boxplots presenting the results of the DVH analysis are shown in Supplementary Material III.D for all three models. In general, minor differences in D_{max} and D_{median} were observed for OARs in DVH analysis for each sequence for the Baseline model. On average, differences were below 0.5% for every sequence and DVH point. Individually, most patients had differences in DVH points ≤1%. Exceptions for FLAIR-based sCT were the pituitary gland (PT14), optic chiasm (PT1), and lens (PT12), with differences ≤2%. PT12 patient had an RT plan with a vast irradiated area, matching the patient's large tumour volume. For this patient, for sCT derived from every MRI sequence, notable dose differences were observed around the body contour on the right half.

3.2.2. Baseline+FLAIR vs Baseline

Training the Baseline+FLAIR model took 46.2 h with the application of early stopping at 450,000 iterations. The MAE obtained on T1w(Gd) (T1w: 65.5 ± 7.7 HU; T1wGd: 64.6 ± 9.3 HU), and T2w (71.2 ± 8.8 HU) images was slightly worse for the Baseline+FLAIR model than the Baseline model (Fig. 3). This difference was statistically significant for T2w and T1w, but not for T1wGd images.

The most notable change in MAE was found on FLAIR images, favouring the Baseline+FLAIR model. Adding FLAIR images to the training data reduced the MAE from 106 ± 20.7 HU to 72.6 ± 10.1 HU ($p < 0.5$). Results for SSIM and PSNR were generally in line with the MAE.

For the Baseline+FLAIR model, $\gamma_{1\%1mm}$-pass rates were >97% for each patient and MRI sequence (Table 3). As for the Baseline model, pass rates $\gamma_{3\%3mm}$, and $\gamma_{2\%2mm}$ were all >99%. For FLAIR images, the Baseline+FLAIR model outperformed the Baseline model in terms of $\gamma_{1\%1mm}$-pass rate: 99.4 ± 0.8% (Baseline+FLAIR model) vs 99.0 ± 1.1% (Baseline model).

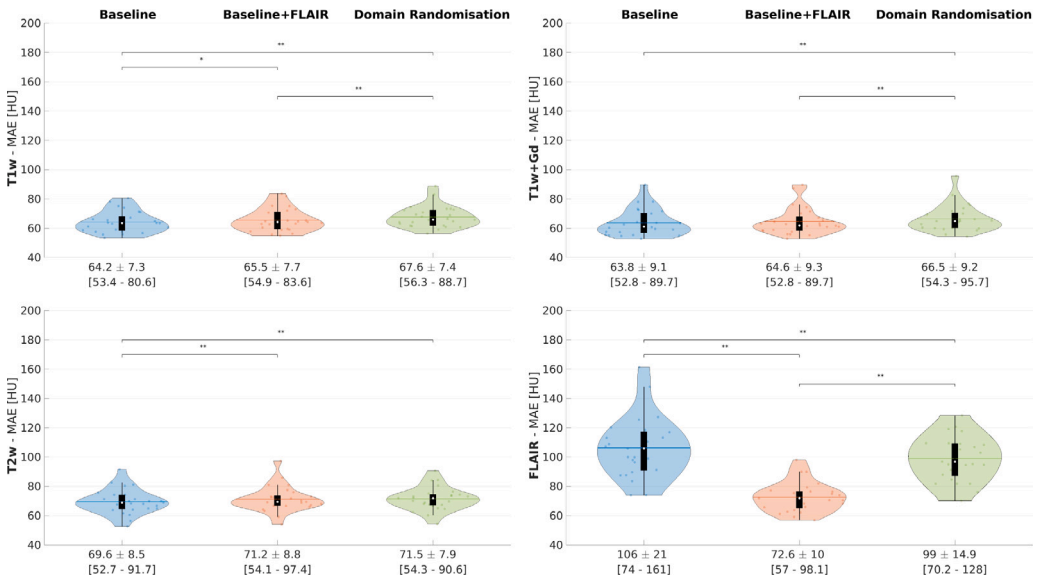

Fig. 3. Violin and box-and-whisker plots for the MAE in the intersection of the body contour of sCT compared to ground truth CT on the test set (n = 25) for sCT generated by the Baseline (blue), Baseline+FLAIR (orange) and Domain Randomisation model (green). Results are presented per sequence: T1w (top left), T1wGd (top right), T2w (bottom left) and FLAIR (bottom right). The black box indicates the interquartile range and median (white circle) with whiskers indicating the range, outliers excluded. The width of the violin indicates the distribution of the data points. The mean values and standard deviations are shown. Statistically significant differences are indicated by * ($p < 0.05$) or ** ($p < 0.001$). Wilcoxon-signed rank tests were used for statistical testing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Dose evaluation ($\gamma_{1\%,1mm}$ and DD) for sCT generated by the Baseline, Baseline+ FLAIR and Domain Randomisation models per MRI sequence. Dose accuracy was assessed through plan re-calculation on sCT_{wf} compared to CT_{wf}. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported.

Metric	Model	Sequence			
		T1w	T1wGd	T2w	FLAIR
$\gamma_{1\%,1mm}$ [%][a]	Baseline	99.4 ± 0.8 [97.1 – 100]	99.4 ± 0.8 [96.9 – 100]	99.4 ± 0.7 [97.3 – 100]	99.0 ± 1.1 [95.4 – 99.9]
	Baseline + FLAIR	99.5 ± 0.7 [97.4 – 100]	99.5 ± 0.7 [97.2 – 100]	99.4 ± 0.7 [97.4 – 100]	99.4 ± 0.8 [97.2 – 100]
	Domain Randomisation	99.4 ± 0.8 [97.0 – 100]	99.4 ± 0.8 [96.9 – 100]	99.3 ± 0.8 [97.2 – 100]	99.2 ± 0.9 [96.6 – 99.9]
DD [%][b]	Baseline	−0.1 ± 0.2 [−0.5 – 0.1]	−0.1 ± 0.2 [−0.5 – 0.1]	−0.1 ± 0.2 [−0.4 – 0.8]	0.4 ± 0.6 [−1.0 – 1.5]
	Baseline + FLAIR	−0.02 ± 0.2 [−0.4 – 0.4]	−0.01 ± 0.2 [−0.7 – 0.5]	0.01 ± 0.3 [−0.4 – 1.1]	0.01 ± 0.4 [−1.4 – 0.7]
	Domain Randomisation	−0.1 ± 0.2 [0.5 – 0.2]	−0.2 ± 0.2 [0.5 – 0.1]	−0.1 ± 0.3 [−0.5 – 0.9]	0.3 ± 0.5 [−0.4 – 1.4]

[a]Calculated in the D > 10% prescribed dose.
[b]Calculated in the D > 90% prescribed dose.

Likewise, the other two pass rates were significantly higher for the Baseline+FLAIR model. Surprisingly, despite the higher MAE obtained in T1w images for the Baseline+FLAIR model versus the Baseline model, a significantly higher $\gamma_{1\%1mm}$-pass rate was obtained for the Baseline+FLAIR model (99.5 ± 0.7% vs 99.4 ± 0.8%). All other differences in pass rates between the two models were not significant.

The absolute DD values obtained per sequence for the Baseline +FLAIR model were smaller than those obtained for the Baseline model ($p < 0.05$), with DD < 1.5% for every patient and seen sequence. For FLAIR images, the number of patients with a DD > 1% was reduced to one (PT2) compared to three for the Baseline model. Similar to what was found for the Baseline model, for PT2, discrepancies between sCT and CT HU values were present near the high-dose region around the surgical intervention.

As for the Baseline model, differences in D_{max} and D_{median} were minor for all DVH points evaluated and all sequences, with average differences <0.5% and DVH point difference <1%, except for the cochlea of PT12 (≤1.5%) probably due to body contour mismatches on the right side.

3.2.3. Domain randomisation vs baselines

The training time for the Domain Randomisation model was 66.4 h (450,000 iterations).

For the seen sequences, the MAE obtained for the Domain Randomisation model was higher (T1w: 67.6 ± 7.4 HU; T1wGd: 66.5 ± 9.2 HU; T2w: 71.5 ± 7.9 HU) than that obtained for the Baseline and Baseline+FLAIR models (Fig. 3). All differences between the Domain Randomisation model and the Baseline model for these three sequences were statistically significant. Likewise, the differences between the Domain Randomisation and the Baseline+FLAIR model were statistically significant for T1w and T1wGd images but not for T2w images. Results for SSIM and PSNR were generally consistent with the MAE.

The MAE obtained for the Domain Randomisation model on FLAIR images (99.0 ± 14.9 HU) was 7 HU lower than that obtained for the Baseline model ($p < 0.05$), a difference which is larger than the increase in MAE obtained for the other sequences (T1w: +3 HU, T1wGd: +3 HU; T2w: +2 HU). Despite this decrease in MAE on FLAIR images obtained through the addition of RC images during network training, the MAE obtained for the Domain Randomisation model was 26 HU

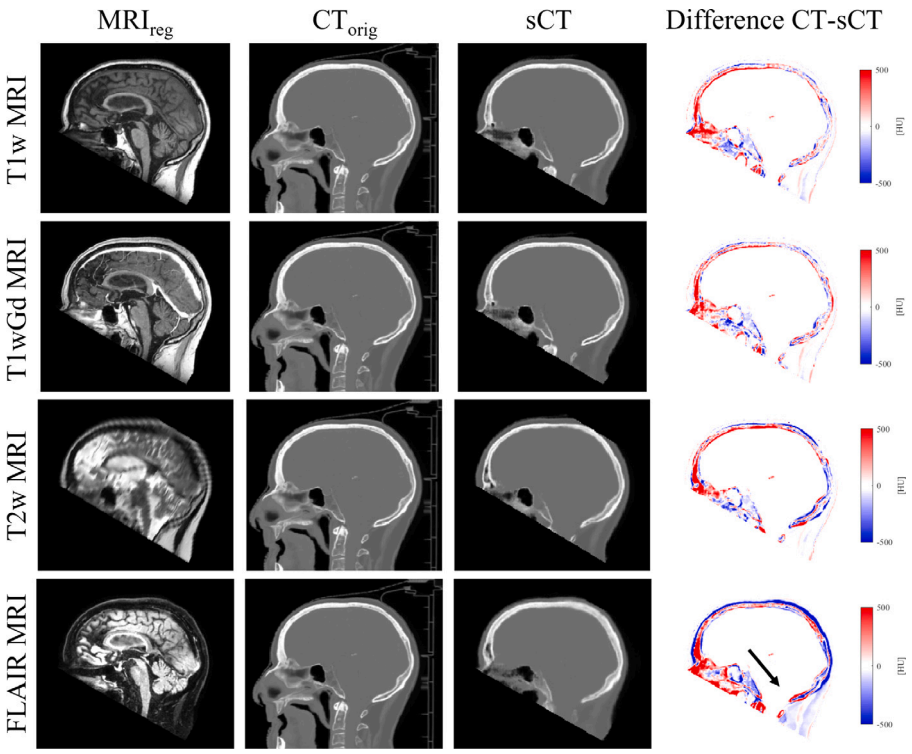

Fig. 4. Results generated by the Domain Randomisation model, for a subject with average performance for T1w, T1wGd, T2w and FLAIR input images (top to bottom). The image shows from left to right: the original MRI, ground truth CT, sCT generated by the Domain Randomisation model, and the difference between the acquired CT and sCT. Typical problematic areas are the nasal cavities and the borders of the skull (bright in the image, with the difference between CT and sCT on the right). For FLAIR specifically, the back of the neck (arrow) is problematic, and the skull is too thick on sCT, represented by the blue colour in the image with the difference between CT and sCT (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

higher than that achievable when adding FLAIR images to the training dataset (Baseline+FLAIR model; $p < 0.05$).

Fig. 4 shows results for an example case generated by the Domain Randomisation model. Typical problematic areas for all sequences are the skull border and the nasal cavities. Similar difficult areas are observed for the Baseline and Baseline+FLAIR model. For FLAIR images, the Domain Randomisation model produced sCTs in which the skull is thicker than the acquired CT, showing a bright blue colour in the difference image (Fig. 4, right). Additionally, the musculature in the back of the neck is typically a problematic area for FLAIR images (arrow). Similar observations are made for the Baseline model. For the Baseline+FLAIR model, the overestimated skull thickness and neck musculature inaccuracies are less prominent.

Overall, a visual inspection of the results generated by the Domain Randomisation model from FLAIR images reveals that the model might be more robust than the Baseline model. Fig. 5 shows three example patients for whom the Baseline model produced artefacts in the sCT (green rectangles). Such artefacts were not observed in the FLAIR-based sCTs produced by the Baseline+FLAIR and Domain Randomisation models. The bottom row shows images of a patient with an oedematous area in the frontal lobe. The area is hypointense on the FLAIR image, leading to an intensity similar to air in the sCT generated by the Baseline model, which translates to a high positive value in the image with the difference between CT and sCT. Results for the Domain Randomisation and Baseline+FLAIR model are less problematic.

Fig. 5 also shows that the Domain Randomisation model better depicts the neck muscles than the baselines. Nevertheless, the smallest differences between CT and sCT in this area are observed in the sCT generated by the Baseline+FLAIR model. The lower differences between sCT and CT in muscle tissue compared to Baseline for the Domain Randomisation model were observed for all patients in the test set. A problem in FLAIR-based sCTs that remains unresolved after applying domain randomisation is the mapping of the skull. Like the Baseline model, the Domain Randomisation model systematically produced sCTs with the skull mapped thicker than on the acquired CT, which is not observed for the Baseline+FLAIR model.

The $\gamma_{1\%1mm}$ for the Domain Randomisation model were >96% for each patient and each MRI sequence (Table 3). For the $\gamma_{3\%3mm}$, $\gamma_{2\%2mm}$, pass rates were all >99%, as for the other two models. Differences in γ-pass rates between the Baseline and Domain Randomisation model were insignificant for the seen sequences. However, for FLAIR images, the Domain Randomisation model outperformed the Baseline model for $\gamma_{1\%1mm}$: 99.2 ± 0.9% vs 99.0 ± 1.1% ($p < 0.05$).

Compared to the Baseline+FLAIR model for FLAIR images, the Domain Randomisation model resulted in significantly lower $\gamma_{1\%1mm}$ (99.2 ± 0.9% vs 99.4 ± 0.8%). Additionally, higher $\gamma_{1\%1mm}$ and $\gamma_{3\%3mm}$-pass rates were obtained for the Baseline+FLAIR model than for the Domain Randomisation model for T1w images ($p < 0.05$).

Differences in DD between the Domain Randomisation model and the Baseline model were not significant. Comparing the DD obtained for the Domain Randomisation and Baseline+FLAIR models resulted in p-values <0.05 for every sequence, with the DD obtained for the Baseline+FLAIR model smaller in absolute terms.

For the Domain Randomisation model, the DD in treatment plans from FLAIR-based sCT was <1.5% and >1% only for three patients (PT13, PT16 and PT18). For PT13 and PT18, we observed differences in the same regions already reported for the Baseline model. For PT16, dose differences in the high-dose region were substantial along the inner border of the skull, in line with the general observation that the MAE along the skull border was comparatively high for sCT generated from FLAIR images.

In general, boxplots for differences in DVH points for OARs reveal slight differences in D_{max} and D_{median} for all DVH points and sequences, with average differences <0.5% as for the other two models. On an individual basis, most patients had differences in DVH points ≤1% for every OAR, except for cochlea of PT6 (<1.5%) and pituitary gland and

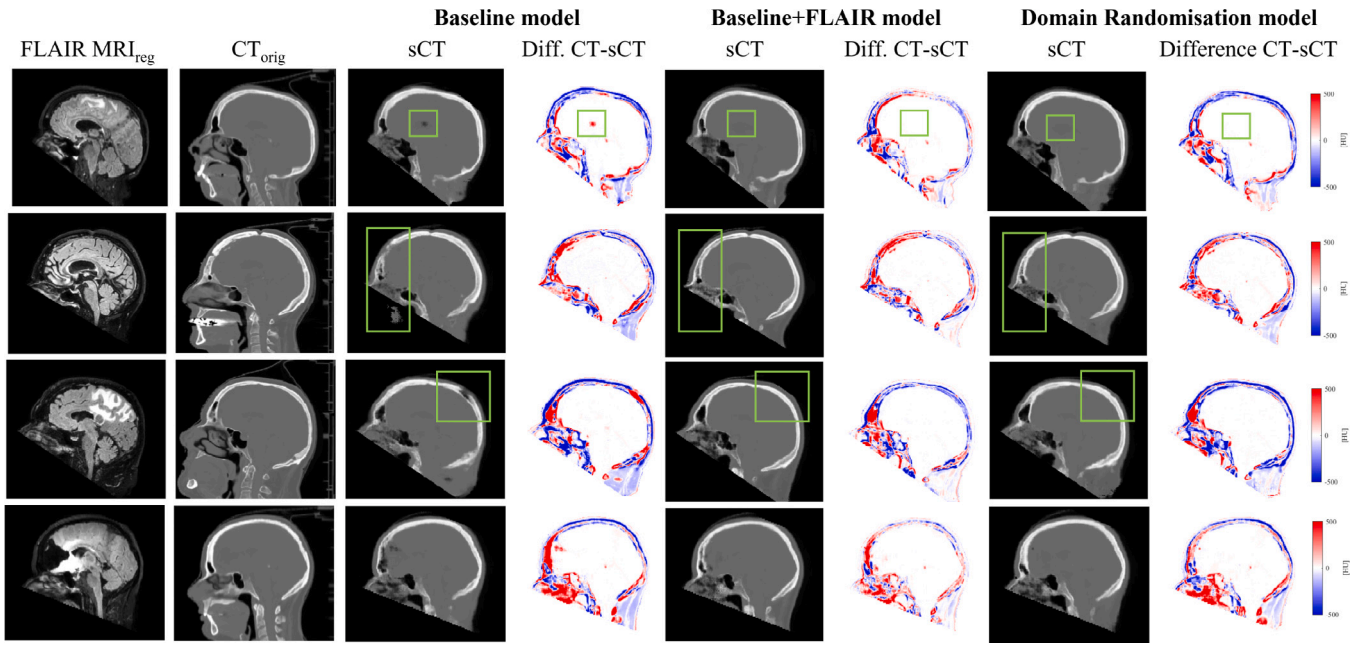

Fig. 5. Results from FLAIR images generated by the three models for four different patients. Images from left to right: original FLAIR MRI, ground truth CT, sCT, and the difference between the acquired CT and sCT for the Baseline model, Baseline+FLAIR model and Domain Randomisation model, respectively. The areas marked with a green rectangle highlight artefacts in sCT produced by the Baseline model that are not present in the sCT generated by the Baseline+FLAIR and Domain Randomisation models. The bottom row shows an example patient with oedema in the frontal lobe. This area is hypointense on the FLAIR image, leading to problems in the sCT generated by the Baseline model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lens of PT12 (<2.5%), which is comparable to the differences observed also for the baselines.

4. Discussion

In this work, we investigated the influence of domain randomisation for MRI-to-sCT generation, considering whether the generalisation to contrasts unseen during training can be increased.

We considered a cGAN model that after optimisation achieved image similarity to CT on par with models presented in the literature (Supplementary Material IV), where the reported range for T1w images is 45.4 HU±8.5 HU [57] to 131 ± 14.3 HU [58] and 44.6 ± 7.5 HU [57] to 89.3 ± 10.3 HU [59] for T1wGd images. The accuracy of dose plans generated from sCT was high for all models and sequences. Considering $\gamma_{3\%3mm}$ and $\gamma_{2\%2mm}$-pass rates for clinical acceptability, dose plans were acceptable for all patients and sequences, even for FLAIR-based sCT generated by a model trained only on a mix of T1w(Gd) and T2w images (Baseline model). Altogether, the number of patients included in our training set (n = 60) is significantly larger than the median number in other studies (n = 33). We believe that our models' performance for the seen sequences is sufficient to explore generalisation, considering the following limitations.

A supervised framework was adopted, requiring a set of well-registered MRI-CT pairs. Poor registration between the CT and MRI in the training dataset is detrimental for models' performance [60]. We recurred to sole rigid registration, resulting in misregistration for the body contours, sinuses, and vertebrae for all the sequences. In the future, non-rigid registrations could be explored to improve the overall model performance or recurring to unpaired training [61].

Hyperparameter tuning was performed using only T1w images, which may not be optimal when mixing other sequences or applying domain randomisation. However, after performing a quick check of the hyperparameters on the RC model, we found that 2–3 HU could decrease MAE through optimisation, which was deemed minimal. Also, the adopted study design without hyperparameter optimisation allows for isolating the effect of domain randomisation.

A critical note should be made about our dose evaluation. Differences in the FOV of the acquired MRI and planning CT led to equal differences between the sCT and planning CT. Water-filling was used to avoid dose differences arising from different FOVs. However, this means that the dose accuracy of the sCT could be overestimated for beams passing through the water-filled area. In this sense, the sCT developed in this work should not be considered for clinical use but can be valuable to shed light on model generalisation.

Overall, we found the performance of a model trained on a mix of T1w(Gd) and T2w images (Baseline) was inferior on FLAIR images compared to performance on the other sequences. We found that domain randomisation generates sCT from FLAIR images with significantly improved image similarity and dose accuracy compared to the Baseline model.

In an additional experiment [50], we found that the benefit of adding RC images to the training data was larger when only T1w acquired MRI were used for network training than a model trained on T1w(Gd) or T1w(Gd) and RC when both T1w(Gd) and T2w images were used (Domain Randomisation model vs Baseline model). Still, having at disposal the unseen sequence was not matched by the domain randomisation method. This means that the MRI sequence that will be clinically used for MRI-only RT should be preferred whenever available. In case such a sequence is not available, domain randomisation can be considered as a method to increase model robustness. It seems that, currently, domain randomisation does not lead to a strong contrast-agnostic method, which contrasts with what was claimed in the original work by Billot et al. [39]. Compared to the original work (segmentation), we applied domain randomisation to a different and more challenging task (image synthesis). Also, Billot et al. [39] did not compare with a statistical test the performance on FLAIR to the performance on other sequences, which complicates judging to what extent their model is contrast-agnostic. Moreover, we speculate that our methods relied on generating RC images from labels, which may result in a loss of within-label structure that may be detrimental to the network considering the image synthesis task. Also, the adopted segmentations were not perfectly aligned with the corresponding ground truth (acquired CT), unlike in [38,39], which might have reduced the effect of the RC

images on network performance in our work. A solution may recur to obtaining through manual segmentation; we considered this procedure too expensive and out of scope for this research. Still, with the chosen study design, we could investigate the impact of domain randomisation. Future studies could clarify whether more accurate or elaborate label maps are more suitable as the basis for RC images, especially if the final goal of domain randomisation is obtaining a contrast-agnostic model.

A second domain randomisation method, based on linear combinations of acquired T1w(Gd) and T2w images, was proposed and tested for the first time. An advantage of this method over RC images is that it requires minimal effort and is easily applicable if multiple sequences are available per patient. However, the contrast produced by this method is less variable than RC, which could explain why this method is not as effective as RC images. Theory and earlier studies suggest that variability beyond what the network will encounter in reality can be beneficial [40,41,62], in line with findings in [39], where synthetic images mimicking specific MRI sequences proved counterproductive. Future work could explore more elaborate domain randomisation methods, i.e., extending LC to non-linear combinations or increasing the number of acquired MRI sequences used for combination. Furthermore, the variability in the RC images could be further improved, e.g., using random elastic deformations or simulation of bias field artefacts as in [38,39]. Another approach could explore using GANs or other DL models to generate synthetic training data, as suggested in, e.g., [63,64].

A relevant question is whether the proposed domain randomisation approach could already be employed clinically to bridge smaller domain gaps than an entirely new sequence, like same-sequence data from a different hospital or changes in the acquisition protocol that might occur over time. Further evaluations on new datasets are needed to investigate whether this is the case.

This work provides the first attempt towards sCT generation from different MRI contrasts for MRI-only RT planning. A clear improvement was found in image similarity for sCT generated from an unseen sequence by Domain Randomisation models compared to baselines. Interestingly, in terms of dose accuracy, our baselines already achieved good results for most patients for the unseen sequence simply by training on a mix of other sequences. The Domain Randomisation model improved the γ-pass rate for the unseen sequence. In contrast, differences with the Baseline model in dose metrics were not statistically significant for the seen sequences, leading us to believe that the small decrease in image similarity obtained for the seen sequences is clinically acceptable. Moreover, the Domain Randomisation model reduced artefacts observed in FLAIR-based sCT comparable to the one observed from the Baseline model. The results indicate that domain randomisation can improve generalisation to unseen sequences for sCT generation. Before clinically implementing the methods described in this work, dose accuracy must be evaluated on MRI acquired with a larger FOV in a clinical setting. Additionally, it is preferable to obtain sCTs with the same voxel size as the acquired data, avoiding the need for resampling. Therefore, further investigations are still required.

The results obtained in this work indicate that domain randomisation might help avoid the need for network re-training if the model is to be used on a sequence unseen during network training. This could be helpful if exceptions need to be made in imaging protocols for specific patients, e.g., due to possible allergic reactions to contrast agents or claustrophobia. On the other hand, each centre should determine whether the performance improvement found in this work is substantial enough to justify the effort associated with implementing domain randomisation, i.e., the need for segmentations, in case alternative sequences are not already available.

5. Conclusion

We investigated the ability of a DL model to generate sCT on unseen sequences accurate for MRI-only radiotherapy. We considered two methods for domain randomisation, showing that adding random contrast images generated from label maps to the training data is more effective than applying random linear combinations of acquired MRI.

Generally, a satisfactory dose accuracy was obtained when training on a mix of acquired sequences, even for the unseen sequence. The adopted domain randomisation method improved dose accuracy and image similarity on this unseen sequence, but could not overperform having at disposal the unseen sequence during training. Domain randomisation can increase model robustness to unseen sequences, reducing the need for model re-training.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejmp.2023.102642.

References

[1] Barton MB, Jacob S, Shafiq J, Wong K, Thompson SR, Hanna TP, et al. Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012. Radiother Oncol 2014;112(1):140–4. http://dx.doi.org/10.1016/j.radonc.2014.03.024.

[2] Seco J, Evans PM. Assessing the effect of electron density in photon dose calculations. Med Phys 2006;33(2):540–52. http://dx.doi.org/10.1118/1.2161407.

[3] Dirix P, Haustermans K, Vandecaveye V. The value of magnetic resonance imaging for radiotherapy planning. Semin Radiat Oncol 2014;24(3):151–9. http://dx.doi.org/10.1016/j.semradonc.2014.02.003.

[4] Jolicoeur M, Racine ML, Trop I, Hathout L, Nguyen D, Derashodian T, et al. Localization of the surgical bed using supine magnetic resonance and computed tomography scan fusion for planification of breast interstitial brachytherapy. Radiother Oncol 2011;100(3):480–4. http://dx.doi.org/10.1016/j.radonc.2011.08.024.

[5] Rasch C, Steenbakkers R, Van Herk M. Target definition in prostate, head, and neck. Semin Radiat Oncol 2005;15(3):136–45. http://dx.doi.org/10.1016/j.semradonc.2005.01.005.

[6] Just M, Rösler HP, Higer HP, Kutzner J, Thelen M. MRI-assisted radiation therapy planning of brain tumors-clinical experiences in 17 patients. Magn Reson Imaging 1991;9(2):173–7. http://dx.doi.org/10.1016/0730-725X(91)90007-9.

[7] Datta N, David R, Gupta R, Lal P. Implications of contrast-enhanced CT-based and MRI-based target volume delineations in radiotherapy treatment planning for brain tumors. J Cancer Res Ther 2008;4(1):9–13. http://dx.doi.org/10.4103/0973-1482.39598.

[8] Jonsson JH, Karlsson MG, Karlsson M, Nyholm T. Treatment planning using MRI data: an analysis of the dose calculation accuracy for different treatment regions. Radiat Oncol 2010;5(1):62. http://dx.doi.org/10.1186/1748-717X-5-62.

[9] Schmidt MA, Payne GS. Radiotherapy planning using MRI. Phys Med Biol 2015;60(22):R323–61. http://dx.doi.org/10.1088/0031-9155/60/22/R323.

[10] Ulin K, Urie MM, Cherlow JM. Results of a multi-institutional benchmark test for cranial CT/MR image registration. 2010/04/08. Int J Radiat Oncol Biol Phys 2010;77(5):1584–9. http://dx.doi.org/10.1016/j.ijrobp.2009.10.017.

[11] Roberson PL, McLaughlin PW, Narayana V, Troyer S, Hixson GV, Kessler ML. Use and uncertainties of mutual information for computed tomography/magnetic resonance (CT/MR) registration post permanent implant of the prostate. Med Phys 2005;32(2):473–82. http://dx.doi.org/10.1118/1.1851920.

[12] Lee YK, Bollet M, Charles-Edwards G, Flower MA, Leach MO, McNair H, et al. Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone. Radiother Oncol 2003;66(2):203–16. http://dx.doi.org/10.1016/S0167-8140(02)00440-1.

[13] Nyholm T, Nyberg M, Karlsson MG, Karlsson M. Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments. Radiat Oncol 2009;4(1):54. http://dx.doi.org/10.1186/1748-717X-4-54.

[14] Kapanen M, Collan J, Beule A, Seppälä T, Saarilahti K, Tenhunen M. Commissioning of MRI-only based treatment planning procedure for external beam radiotherapy of prostate. Magn Reson Med 2013;70(1):127–35. http://dx.doi.org/10.1002/mrm.24459.

[15] Khong P, Ringertz H, Donoghue V, Frush D, Rehani M, Appelgate K, et al. ICRP publication 121: Radiological protection in paediatric diagnostic and interventional radiology. Ann ICRP 2013;42:1–63. http://dx.doi.org/10.1016/j.icrp.2012.10.001.

[16] Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. Radiat Oncol 2017;12(1):28. http://dx.doi.org/10.1186/s13014-016-0747-y.

[17] Owrangi AM, Greer PB, Glide-Hurst CK. MRI-only treatment planning: Benefits and challenges. Phys Med Biol 2018;63(5). http://dx.doi.org/10.1088/1361-6560/aaaca4.

[18] Karlsson M, Karlsson MG, Nyholm T, Amies C, Zackrisson B. Dedicated magnetic resonance imaging in the radiotherapy clinic. Int J Radiat Oncol Biol Phys 2009;74(2):644–51. http://dx.doi.org/10.1016/j.ijrobp.2009.01.065.

[19] Devic S. MRI simulation for radiotherapy treatment planning. Med Phys 2012;39(11):6701–11. http://dx.doi.org/10.1118/1.4758068.

[20] Lagendijk JJW, Raaymakers BW, Raaijmakers AJE, Overweg J, Brown KJ, Kerkhof EM, et al. MRI/linac integration. Radiother Oncol 2008;86(1):25–9. http://dx.doi.org/10.1016/j.radonc.2007.10.034.

[21] Mutic S, Dempsey JF. The ViewRay system: Magnetic resonance–guided and controlled radiotherapy. Semin Radiat Oncol 2014;24(3):196–9. http://dx.doi.org/10.1016/j.semradonc.2014.02.008.

[22] Raaymakers BW, Jürgenliemk-Schulz IM, Bol GH, Glitzner M, Kotte ANTJ, van Asselen B, et al. First patients treated with a 1.5 T MRI-Linac: clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment. Phys Med Biol 2017;62(23):L41–50. http://dx.doi.org/10.1088/1361-6560/aa9517.

[23] Hall WA, Paulson ES, van der Heide UA, Fuller CD, Raaymakers BW, Lagendijk JJW, et al. The transformation of radiation oncology using real-time magnetic resonance guidance: A review. Eur J Cancer 2019;122:42–52. http://dx.doi.org/10.1016/j.ejca.2019.07.021.

[24] Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning-based synthetic-CT generation in radiotherapy and PET: a review. Med Phys 2021;48(11):6537–66. http://dx.doi.org/10.1002/mp.15150.

[25] Kerkmeijer LGW, Maspero M, Meijer GJ, van der Voort van Zyp JRN, de Boer HCJ, van den Berg CAT. Magnetic resonance imaging only workflow for radiotherapy simulation and planning in prostate cancer. Clin Oncol 2018;30(11):692–701. http://dx.doi.org/10.1016/j.clon.2018.08.009.

[26] Boulanger M, Nunes JC, Chourak H, Largent A, Tahri S, Acosta O, et al. Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. Phys Med 2021;89:265–81.

[27] Klages P, Benslimane I, Riyahi S, Jiang J, Hunt M, Deasy JO, et al. Patch-based generative adversarial neural network models for head and neck MR-only planning. 2019/12/25. Med Phys 2020;47(2):626–42. http://dx.doi.org/10.1002/mp.13927.

[28] O'Connor LM, Choi JH, Dowling JA, Warren-Forward H, Martin J, Greer PB. Comparison of synthetic computed tomography generation methods, incorporating male and female anatomical differences, for magnetic resonance imaging-only definitive pelvic radiotherapy. Front Oncol 2022;12(822687). http://dx.doi.org/10.3389/fonc.2022.822687.

[29] Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research (PMLR), vol. 97, 2019, p. 5389–400, URL: https://proceedings.mlr.press/v97/recht19a.html.

[30] Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012;45(1):521–30. http://dx.doi.org/10.1016/j.patcog.2011.06.019.

[31] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59. http://dx.doi.org/10.1109/TKDE.2009.191.

[32] Brou Boni KND, Klein J, Gulyban A, Reynaert N, Pasquier D. Improving generalization in MR-to-CT synthesis in radiotherapy by using an augmented cycle generative adversarial network with unpaired data. Med Phys 2021;48(6):3003–10. http://dx.doi.org/10.1002/mp.14866.

[33] Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Phys Med Biol 2022;67(11):11TR01. http://dx.doi.org/10.1088/1361-6560/ac678a.

[34] Fiorino C, Jeraj R, Clark CH, Garibaldi C, Georg D, Muren L, et al. Grand challenges for medical physics in radiation oncology. Radiother Oncol 2020;153:7–14. http://dx.doi.org/10.1016/j.radonc.2020.10.001.

[35] Bosmans H, Zanca F, Gelaude F. Procurement, commissioning and QA of AI based solutions: An MPE's perspective on introducing AI in clinical practice. Phys Med 2021;83:257–63. http://dx.doi.org/10.1016/j.ejmp.2021.04.006.

[36] Li W, Kazemifar S, Bai T, Nguyen D, Weng Y, Li Y, et al. Synthesizing CT images from MR images with deep learning: Model generalization for different datasets through transfer learning. Biomed Phys Eng Express 2021;7(2):025020. http://dx.doi.org/10.1088/2057-1976/abe3a7.

[37] Zimmermann L, Knäusl B, Stock M, Lütgendorf-Caucig C, Georg D, Kuess P. An MRI sequence independent convolutional neural network for synthetic head CT generation in proton therapy. Z Med Phys 2021. http://dx.doi.org/10.1016/j.zemedi.2021.10.003.

[38] Billot B, Greve DN, Van Leemput K, Fischl B, Iglesias JE, Dalca A. A learning strategy for contrast-agnostic MRI segmentation. In: Arbel T, Ben Ayed I, de Bruijne M, Descoteaux M, Lombaert H, Pal C, editors. Proceedings of the third conference on medical imaging with deep learning. Proceedings of machine learning research (PMLR), vol. 121, 2020, p. 75–93, URL: https://proceedings.mlr.press/v121/billot20a.html.

[39] Billot B, Greve DN, Puonti O, Thielscher A, Van Leemput K, Fischl B, et al. SynthSeg: Domain randomisation for segmentation of brain scans of any contrast and resolution. 2021, URL: http://arxiv.org/abs/2107.09559.

[40] Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems. 2017, p. 23–30. http://dx.doi.org/10.1109/IROS.2017.8202133.

[41] Bengio Y, Bastien F, Bergeron A, Boulanger-Lewandowski N, Breuel T, Chherawala Y, et al. Deep learners benefit more from out-of-distribution examples. In: Gordon G, Dunson D, Dudík M, editors. Proceedings of the fourteenth international conference on artificial intelligence and statistics. Proceedings of machine learning research (PMLR), vol. 15, 2011, p. 164–72, URL: https://proceedings.mlr.press/v15/bengio11b.html.

[42] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: A toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196–205. http://dx.doi.org/10.1109/TMI.2009.2035616.

[43] Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform 2014;7:50. http://dx.doi.org/10.3389/fninf.2013.00050.

[44] Maspero M, Bentvelzen LG, Savenije MHF, Guerreiro F, Seravalli E, Janssens GO, et al. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. Radiother Oncol 2020;153:197–204. http://dx.doi.org/10.1016/j.radonc.2020.09.029.

[45] Hissoiny S, Ozell B, Bouchard H, Després P. GPUMCD: A new GPU-oriented Monte Carlo dose calculation platform. Med Phys 2011;38:754–64. http://dx.doi.org/10.1118/1.3539725.

[46] Korsholm ME, Waring LW, Edmund JM. A criterion for the reliable use of MRI-only radiotherapy. Radiat Oncol 2014;9(1):16. http://dx.doi.org/10.1186/1748-717X-9-16.

[47] Low DA, Harms WB, Mutic S, Purdy J. A technique for the quantitative evaluation of dose distributions. Med Phys 1998;25(5):656–61. http://dx.doi.org/10.1118/1.598248.

[48] Heilemann G, Poppe B, Laub WU. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. Med Phys 2013;40(3):031702. http://dx.doi.org/10.1118/1.4789580.

[49] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE conference on computer vision and pattern recognition. 2017, p. 5967–76. http://dx.doi.org/10.1109/CVPR.2017.632.

[50] Nijskens L. Contrast generalisation in deep learning-based brain MRI-to-CT synthesis [MSc], University of Twente; 2022, Available at https://essay.utwente.nl/91108/.

[51] Hadzic I, Pai S, Chinmay R, Teuwen J. ganslate-team/ganslate: Initial public release. Zenodo; 2021, http://dx.doi.org/10.5281/zenodo.5494572.

[52] Kingma D, Ba J. ADAM: A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations. 2015, p. 1–15, URL: https://arxiv.org/abs/1412.6980.

[53] Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 2020;219:117012. http://dx.doi.org/10.1016/j.neuroimage.2020.117012.

[54] Savenije MHF, Maspero M, Sikkes GG, van der Voort van Zyp JRN, Kotte ANTJ, Bol GH, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiat Oncol 2020;15(1):104. http://dx.doi.org/10.1186/s13014-020-01528-0.

[55] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78. http://dx.doi.org/10.1016/j.media.2016.10.004.

[56] Pérez-García F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 2021;208:106236. http://dx.doi.org/10.1016/j.cmpb.2021.106236.

[57] Massa H, Johnson J, McMillan A. Comparison of deep learning synthesis of synthetic CTs using clinical MRI inputs. Phys Med Biol 2020;65. http://dx.doi.org/10.1088/1361-6560/abc5cb.

[58] Irmak S, Zimmermann L, Georg D, Kuess P, Lechner W. Cone beam CT based validation of neural network generated synthetic CTs for radiotherapy in the head region. Med Phys 2021;48(8):4560–71. http://dx.doi.org/10.1002/mp.14987.

[59] Emami H, Dong M, Nejad-Davarani S, Glide-Hurst C. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. Med Phys 2018;45. http://dx.doi.org/10.1002/mp.13047.

[60] Florkow MC, Zijlstra F, Kerkmeijer LGW, Maspero M, Van Den Berg CAT, Van Stralen M, et al. The impact of MRI-CT registration errors on deep learning-based synthetic CT generation. In: Medical imaging 2019: Image processing, Proceedings of SPIE, vol. 10949. 2019, http://dx.doi.org/10.1117/12.2512747.

[61] Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT synthesis using unpaired data. In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL, editors. Simulation and synthesis in medical imaging. Lecture notes in computer science (LNCS), vol. 10557, 2017, p. 14–23. http://dx.doi.org/10.1007/978-3-319-68127-6_2.

[62] Tremblay J, Prakash A, Acuna D, Brophy M, Jampani V, Anil C, et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops. 2018, p. 1082–10828.

[63] Nogues FC, Huie A, Dasgupta S. Object detection using domain randomization and generative adversarial refinement of synthetic images. 2018, http://dx.doi.org/10.48550/ARXIV.1805.11778, arXiv, URL: https://arxiv.org/abs/1805.11778.

[64] Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci Rep 2019;9:16884. http://dx.doi.org/10.1038/s41598-019-52737-x.