Comparison of deep learning-based and patch-based methods for pseudo-CT generation in MRI-based prostate dose planning

Axel Largent, Anais Barateau, Jean-Claude Nunes, Eugenia Mylona, Joel Castelli, Caroline Lafond, Peter B Greer, Jason A Dowling, John Baxter, Hervé Saint-Jalmes, et al.

Version of Record: https://www.sciencedirect.com/science/article/pii/S0360301619337356
Manuscript_9063cc71f94ebef26c92b9f048c868fd

Comparison of deep learning-based and patch-based methods for pseudo-CT generation in MRI-based prostate dose planning

Axel Largent[1], PhD; Anaïs Barateau[1], MSc; Jean-Claude Nunes[1], PhD; Eugenia Mylona[1], MSc; Joël Castelli[1], MD; Caroline Lafond[1], PhD; Peter B. Greer[2, 3], PhD; Jason A. Dowling[4], PhD; John Baxter[1], PhD; Hervé Saint-Jalmes[1], PhD; Oscar Acosta[1], PhD; Renaud de Crevoisier[1], MD

1. Univ Rennes, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
2. School of Mathematical and Physical Sciences University of Newcastle/Newcastle/Australia
3. Department of Radiation Oncology, Calvary Mater, Newcastle, Australia
4. CSIRO Australian e-Health Research Centre, Herston/Queensland/Australia

ACKNOWLEDGMENTS

The authors thank Eugenia Mylona for her contribution to statistical analyses, especially her strong expertise in permutation tests. This work was supported by Cancer Council New South Wales research grant RG11-05, the Prostate Cancer Foundation of Australia (Movember Young Investigator Grant YI2011), and Cure Cancer Australia.

Axel Largent was responsible for statistical analysis.

Conflicts of interest: None.

Comparison of deep learning-based and patch-based methods for pseudo-CT generation in MRI-based prostate dose planning

Abstract

Purpose

Deep learning methods (DLMs) have recently been proposed to generate pseudo-CT (pCT) for MRI-based dose planning. This study aims to evaluate and compare DLMs (U-Net and generative adversarial network (GAN)) using various loss functions (L2, single-scale perceptual loss (PL), multiscale PL, weighted multiscale PL), and a patch-based method (PBM).

Materials and Methods

Thirty-nine patients received a VMAT for prostate cancer (78 Gy). T_2-weighted MRIs were acquired in addition to planning CTs. The pCTs were generated from the MRIs using seven configurations: four GANs (L2, single-scale PL, multiscale PL, weighted multiscale PL), two U-Net (L2 and single-scale PL), and the PBM. The imaging endpoints were mean absolute error (MAE) and mean error (ME), in Hounsfield units (HU), between the reference CT (CT_{ref}) and the pCT. Dose uncertainties were quantified as mean absolute differences between the DVHs calculated from the CT_{ref} and pCT obtained by each method. 3D gamma indexes were analyzed.

Results

Considering the image uncertainties in the whole pelvis, GAN L2 and U-Net L2 showed the lowest MAE (≤34.4 HU). The ME were not different than 0 ($p \leq 0.05$). The PBM provided the highest uncertainties. Very few DVH points differed when comparing GAN L2 or U-Net L2 DVHs and CT_{ref} DVHs ($p \leq 0.05$). Their dose uncertainties were: ≤0.6% for the prostate PTV $V_{95\%}$, ≤0.5% for the rectum V_{70Gy}, and ≤0.1% for the bladder V_{50Gy}. The PBM, U-Net PL and GAN PL presented the highest systematic dose uncertainties. The gamma passrates were >99% for all DLMs. The mean calculation time to generate one pCT was 15 s for the DLMs and 62 min for the PBM.

Conclusion

Generating pCT for MRI dose planning with DLMs and PBM provided low dose uncertainties. In particular, the GAN L2 and U-Net L2 provided the lowest dose uncertainties together with a low computation time.

Keywords: pseudo-CT generation; MRI-only radiotherapy; deep learning; dose calculation; prostate cancer

INTRODUCTION

MRI is clearly superior to CT for organ delineation and could therefore improve tumor targeting in dose planning (1). However, MRI does not provide electron density information that is necessary for dose calculation. To overcome this issue, several methods have been developed to generate pseudo-CTs (pCTs) for MRI-based dose planning (2, 3). These methods can be divided into four categories: bulk density methods (BDM) (4–8), probabilistic methods (9), atlas-based methods (ABM) (10–17), and more recently machine learning methods such as patch-based methods (PBM) including random forest modeling (18–22) and deep learning methods (DLM) (23–29). The BDMs assign homogeneous densities to the volumes of interest (VOIs) that are manually delineated from the patient's MRI. Probabilistic methods use the probability density function to determine the Hounsfield Unit (HU) in each voxel. The ABMs involve complex non-rigid registrations of CT-MRI atlases with the patient's MRI, followed by a CT fusion step to obtain the pCT. The PBMs select the k closest CT patches from a training cohort for a given MRI patch from the patient. The selected CT patches are then fused to generate the corresponding pCT.

Deep learning methods (DLMs) enable the computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction (30). Deep learning has recently been introduced in radiotherapy for multiple applications, such as image segmentation, image processing and reconstruction, image registration, treatment planning, and radiomics (31–37). DLMs have been more recently proposed for pCT generation from MRI (38–43). They are particularly appealing owing to their fast computation time. These methods model relations between the HU values of the CTs and the intensities of the MRIs by training neural networks. Once the optimal network parameters are estimated, the model can be finally applied to a test patient MRI to generate its corresponding pCT. One of the first DLMs for pCT generation from MRI was based on the U-Net architecture (U-Net DLM) (23). More recently, DLMs that use a generative adversarial network (GAN DLM) architecture have also been proposed (24, 25, 27, 29, 44) (Fig. 1), with the theoretical advantage of GAN compared to U-Net to provide more realistic pCTs by obtaining an adversarial feedback from a discriminator network. While GAN and U-Net DLMs provide promising preliminary results, they use most often a standard loss function (L2 and L1 norms) which may also produce blurring and loss of details (29). Perceptual loss could overcome this issue by mimicking human visual perception using similar features (such as multiscale features) but it has never been investigated in this pCT generation application (45–47). Network hyperparameters such as layer level, the number

and weight associated to each level (for perceptual loss), and the discriminator weight compared to the generator weight can also affect the image accuracy. Overall, all these DLM configurations lack a thorough dose evaluation for pCT generation from MRI.

We previously showed that PBM provided lower imaging and dose uncertainties in the pelvis compared with ABM and BDM (20). PBM was found to be faster than ABM. In another study, the U-Net DLM with L2 loss function has been shown to provide better imaging results than the ABM, similar dosimetric results as the ABM, and fewer uncertainties than BDM (48). However, even though the PBMs and DLMs can be considered the most suitable methods for MRI-based dose planning, they have never been compared. Finally, U-Net and GAN DLMs have never been dosimetrically compared in the literature.
This study aims to evaluate and compare the U-Net and GAN DLMs using various hyperparameters and loss functions (L2, single-scale PL, multiscale PL, weighted multiscale PL) as well as PBM, for prostate cancer MRI-only dose planning.

MATERIALS AND METHODS

Thirty-nine patients received a volumetric modulated arc therapy (VMAT) for localized prostate cancer. The ethics approval for the study protocol was provided by the local area health ethics committee and informed consent was obtained from all patients (10). The study follows the same workflow described in our previous study (20).

Image acquisition

Patients had both an initial CT ($CT_{initial}$) and 3T MR imaging (MRI) in the treatment position (Appendix 1) (20). The CT scans were acquired with a GE LightSpeedRT large-bore scanner or a Toshiba Aquilion. The MRI was acquired with a 3T Siemens Skyra MRI scanner. For MRI acquisition, 3D T_2-weighted SPACE sequences were considered with the following parameters: TE = 102 ms, TR = 1200 ms, flip angle = 35°, field-of-view = 430 × 430 × 200 mm^3, and voxel size = 1.6 mm^3.

MRI preprocessing and intra-patient CT to MRI registration

The T_2-weighted images were preprocessed for normalization and correction of image nonuniformity (Appendix 2) (10, 12). Even if the delay between the acquisition of $CT_{initial}$ and MRI was kept as short as possible, the patient's anatomy could still be different between acquisitions. To minimize these pelvic anatomy variations between CT and MRI (10), each $CT_{initial}$ was registered to its corresponding MRI by using a rigid registration (49) followed by a non-rigid registration (50). This registered CT was considered as the reference (CT_{ref}).

For all pCT generation methods, the entire cohort (39 patients) was randomly split three times with non-repeated patients between training (N = 25) and validation cohorts. For validation, the model was trained independently on each of the three different training cohorts. The patients in the validation cohorts were all different (14 + 14 + 11 patients, respectively). Thus, the number of patients in the training/validation cohorts were: 25/14, 25/14 and 25/11.

Patch-based method for pseudo-CT generation

The PBM is detailed in (20) and Appendix 3. To summarize, this method can be divided into the four following steps.

(1) An inter-patient rigid and affine group-wise registration was performed to match all pre-processed MR images into the same coordinate system. Then, the obtained transformations were applied to the corresponding CT images to propagate them into the same coordinate system.

(2) A feature extraction step was performed to obtain spatial, textural, and gradient information from the registered MRI, followed by patch partitioning with overlap (51). The selected features were the multi-scale MR intensities, Shannon entropy, and the norm of the gradient (51). The patch partitioning was conducted on each feature image and the related CT image. The Cartesian coordinates of the centered location of the patches were used as the spatial information.

(3) An approximate nearest neighbor (ANN) search model (52) was generated to select the training patches closest to the target MRI patches. Several randomized KD-trees were trained on the full training feature patch set. These KD-trees aimed to organize the feature patches in a data structure, thereby performing the nearest neighbor search more efficiently. The feature patches from the target MRI were iteratively given as the input of the randomized KD-trees.

Ten feature patches (from the training cohort) closest to the target feature patches were then successively selected. After each iteration, only the CT patches related to the ten closest feature patches were stored.

(4) A multipoint-wise aggregation scheme was conducted to generate the pCT patches. For each target feature patch centered at a location v, only the closest related CT patches near v were fused by weighted means. The weights were obtained by computing the normalized Euclidian distances between the target feature patch and the closest feature patches. The weighted mean was used to estimate the pCT HU value at location v.

The PBM was implemented in C++ using the Insight ToolKit library (53). The training computation time was approximately 24 h (without GPU and cluster architecture).

Deep learning methods for pseudo-CT generation

Fig. 1 depicts the overall workflow of the compared deep learning methods with distinct implemented loss functions. As illustrated, two different networks (U-NET and GAN) trained with different loss functions constituted a set of six training strategies: i) U-Net with L2 loss (U-Net L2), ii) U-Net with single-scale perceptual loss (U-Net PL), iii) GAN with L2 loss (GAN L2), iv) GAN with single-scale perceptual loss (GAN PL), v) GAN with multi-scale perceptual loss (GAN MPL), and vi) GAN with weighted multi-scale perceptual loss (GAN WMPL).

U-Net deep learning method

The U-Net DLM was implemented based upon a 2D architecture similar to the one proposed by Han (23). This architecture was composed of two networks called encoding and decoding parts. The encoding part aimed to extract the multi-scale features from the target MRI. This network was composed of 12 convolutional layers, followed by batch normalization and ReLu activation functions (54). The filter numbers of these layers were 64, 64, 128, 128, 256, 256, 256, 512, 512, 512, 512, and 512, and the filter size was 3×3 (stride = 1). To obtained multi-scale information, some of the features were downsampled using convolutional layers with a filter size of 2×2 and stride = 2.

The decoding part aimed to gradually reconstruct the pCT using the features computed in the encoding part. This network was a mirror version of the encoding part. For feature up-sampling, transposed 2D convolutional layers were used with a filter size of 2 × 2 and stride = 2. To obtain the pCT, the last layer of the decoding part was a convolution layer with one filter (size = 1×1).

One of the differences between our U-Net DLM and the one proposed by Han (23) is for feature map down-sampling and up-sampling. We used 2D convolutional filters (with stride = 2x2) and 2D transpose convolutional filters, instead of max pooling and up pooling. The advantage of using these convolutional filters is that their related weights can be optimized during the training process, allowing computation of new features for better data representation. The max pooling is a fixed operation where no new feature is computed. Additionally, we added batch normalization after some convolutional layers to improve the convergence of the loss function during the gradient descent. Finally, the number of convolutional layers linking the encoding and decoding parts was decreased. The aim of this change was to reduce the blur effect in pCTs, arising when applying too many convolution filters to the low resolution feature maps.

As shown in Fig. 1, to train our U-Net DLM, two different loss functions were implemented: L2 loss (23, 29) and single-scale perceptual loss (45). The L1 loss function was not considered because it was used as an evaluation metric (see imaging endpoints section below). The L2 loss aimed to minimize the differences between the CT and pCT voxels. This loss function was defined as:

$$L_{U\text{-}Net}(I, C) = \|C - U\text{-}Net(I)\|_2^2$$

Where I is the MRI, C is the corresponding CT, $U\text{-}Net(I)$ is the pCT generated by the U-Net, and $\| \|_2^2$ is the L2 norm.

The single-scale perceptual loss mimics the human visual system to compare CT and pCT images using similar features as opposed to only the intensities (24, 45). The VGG16 network was pretrained from the ImageNet data set, available in Keras (55), and used to compute the features inside the CT and pCT images.. The choice of VGG16 was justified because this network is often used for perceptual loss computation in the literature and appears relevant for different tasks (image deblurring, super-resolution, etc.) (45, 55). The perceptual loss function was defined as:

$$L_{U\text{-}Net}(I, C) = \left\|VGG(C) - VGG\left(U\text{-}Net(I)\right)\right\|_2^2$$

where VGG is the output of the 7th VGG16 convolutional layer. The choice 7th VGG layer is justified in Appendix 4.1.

Generative Adversarial Network (GAN) deep learning method

The GAN DLM architecture was composed of two networks: a generator (G) and a discriminator (D), which were trained in competition with each other and illustrated in Fig. 1.

Generator network

The generator network aimed to provide pCTs from the patient MRIs. The generator network used a 2D architecture identical to the previously described U-Net DLM. Besides the previously defined L2 (56) and single-scale perceptual loss functions, two multi-scale versions of perceptual losses were implemented, including a weighted multi-scale implementation.

The evenly weighted multi-scale perceptual loss aimed to first compute the L2 norm between the CT and pCTs feature for some VGG layers. These layers correspond to each scale change in the VGG architecture. Then, the obtained L2 norms integrated in the perceptual loss were averaged considering the multi-scale information of each layer (Appendix 4.2). This multi-scale perceptual loss was described as:

$$L_G(I, C) = \frac{1}{card(S)} \sum_{i \in S} \|VGG_i(C) - VGG_i(G(I))\|_2^2$$

Where $S = \{2, 5, 7, 10, 13\}$, I is the MRI, C is the corresponding CT, $G(I)$ is the pCT produced by the generator, VGG_i is the i^{th} VGG16 convolutional layer, and $\| \|_2^2$ is the L2 norm.

The weighted version of multi-scale perceptual loss follows the same principle as the loss described previously. However, the L2 norms obtained from the VGG layers were weighted to give more importance to the layers yielding the lowest MAE (Appendix 4.2). The weighted multi-scale perceptual loss was described as follows:

$$L_G(I, C) = \frac{1}{card(S)} \sum_{i \in S} w_i \|VGG_i(C) - VGG_i(G(I))\|_2^2$$

Where $w_i = e^{-\left(MAE_i(C, G(I))\right)}$ with MAE_i is the mean absolute error between CTs and pCTs generated by the GAN using the i^{th} VGG16 convolutional layer for perceptual loss computation. The considered MAEs were computed inside the whole pelvis.

Discriminator network

The discriminator network aimed to classify the generated pCT image as real or fake CT. Thus, the output of this network is a probability value ranging between 0 and 1 depending on whether the generated pCT seems to be fake or real, respectively. The architecture was composed of six convolutional layers and one fully connected layer. Each convolutional layer was followed by batch normalization and Leaky-ReLu activation functions. The number of filters for these layers were 8, 16, 32, 64, 64 and 64. The filter size was 3 × 3 (stride = 2) for the first four layers and 1 × 1 (stride = 1) for the remaining layers. The fully connected layer had one filter followed by a sigmoid activation function.

The loss function of the discriminator was a binary cross entropy (29, 45, 57) defined as: $L_D(G(I), C) = -\sum_{i=1}^{n} C_i log(G(I)_i) + (1 - C_i) \log(1 - G(I)_i)$, where $G(I)$ is the pCT computed by the generator from the target MRI I, C is the corresponding CT, and n is the number of voxels inside the C and I images.

The generator and discriminator losses were combined to form the following adversarial loss: $L_{adversarial}(I, C) = \lambda_1 L_D(I, C) + \lambda_2 L_G(I, C)$, where I is the MRI, C is the corresponding CT, $L_D(I, C)$ is the discriminator loss, $L_G(I, C)$ is the generator loss, and λ_1 and λ_2 are the weights for the discriminator and generator losses, respectively. The discriminator was first trained using the discriminator loss, followed by, the generator training using the fully adversarial loss. These training steps were performed iteratively and stopped when the discriminator could not accurately determine if the pCTs provided by the generator looked like true or false CTs.

Training of the U-Net and GAN methods

The U-Net and GAN DLMs were trained using anatomically paired data: axial 2D slices of the training CT and MR images (3600 slices). Data augmentation was performed to increase the size of the training cohort. It was conducted by randomly applying affine registrations on the slices (translated by -5% to 5% per axis, rotated by -10° to +10°, sheared by -10° to 10°). A mini-batch size of 5 slices and 300 epochs was considered. The choice of mini-batch size is detailed in Appendix 4.3. The network parameters were optimized using the Adam algorithm (58). The parameters of this algorithm parameters were: $\alpha = 1 \times 10^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.9$. For the GAN, the weights of the discriminator and generator loss functions were: $\lambda_1 = 5$ and $\lambda_2 = 1$, respectively. The convergence of GAN with perceptual loss (generator and discriminator) is presented in Appendix 4.4.

The U-Net and GAN DLMs were implemented in Python using Keras (59). The training computation time for the networks was approximately 24 h each with a GPU Nvidia GTX 1070 TI 8 GB.

The stochastic effect on the training of each pCT generation method (U-Net, GAN, and PBM) was assessed by repeating three pCT generations (training and validation) for each group (25/14, 25/14 and 25/11) and for each method (Appendix 5).

Delineation and dose calculation on reference CT and pseudo-CT

Organ delineation was performed on CT_{ref}, by a senior oncologist, in agreement with the GETUG/RECORAD group recommendation (Appendix 6) (60). The contours were rigidly propagated from CT_{ref} to pCT.

A VMAT was planned on the CT_{ref} images with the Pinnacle v.9.10 (Philips) treatment planning system for prostate and seminal vesicles. The collapsed cone convolution algorithm was used for dose calculation. A sequential treatment was delivered with a total dose of 50 Gy to the prostate and seminal vesicles, followed by a boost of 28 Gy in the prostate (at 2 Gy per fraction). GETUG dose–volume constraints were applied to the organs-at-risk (OARs) (Appendix 6) (60). The beam parameters used to compute the dose from CT_{ref} were used to calculate the dose from pCT.

Endpoints and statistical analyses

Imaging and dosimetric endpoints were considered for the 39 patients in a cross validation, using the seven pCT generation configurations: PBM, U-Net with L2 loss (U-Net L2), U-Net with single-scale perceptual loss (U-Net PL), GAN with L2 loss (GAN L2), GAN with single-scale perceptual loss (GAN PL), GAN with multi-scale perceptual loss (GAN MPL), and GAN with weighted multi-scale perceptual loss (GAN WMPL).

Imaging endpoints

To compare the imaging accuracy of different pCT generation methods, a voxel-wise comparison of the HU between CT_{ref} and pCT was performed. To accomplish this, the mean absolute error (MAE) and the mean error (ME) were calculated between the CT_{ref} and pCT obtained from the seven configurations. These endpoints were defined as: $MAE = \frac{1}{n}\sum_{i=1}^{n}|HU_{CTref}(i) - HU_{pCT}(i)|$ and $ME = \frac{1}{n}\sum_{i=1}^{n} HU_{CTref}(i) - HU_{pCT}(i)$. They were calculated in the entire body, soft tissues (prostate, rectum and bladder) and pelvic bones (femoral heads). Additional Table 1 lists the mean HU values of the CT_{ref} inside each VOI.

Dosimetric endpoints

The accuracy of the methods was first evaluated by computing the dose uncertainty (MAE) and systematic dose uncertainty (ME). The dose uncertainty was defined by the differences in mean absolute values across dose volume histograms (DVHs) calculated from the dose on the CT_{ref} and the pCTs. The systematic dose uncertainty was computed as the mean DVH differences between the CT_{ref} and pCT. These uncertainties were reported for the RTOG/GETUG reference DVH points (60, 61) and the entire DVH of the VOI (PTV prostate, bladder, rectum and femoral heads). The DVH bin size was 5 cGy. The mean dose (D_{mean}) was also considered. A spatial dose evaluation was finally conducted by performing 3D gamma analyses (local, 1%/1 mm, dose thresholds 10% and 30%) using the dose distributions from CT_{ref} and pCTs.

Statistical analysis

Wilcoxon signed-rank tests were performed to compare the endpoints. For the MAE (image and dose), these tests were used to compare the lowest MAE among all the methods to the MAE of each other method, and also to compare MAE of the GAN PL method to the MAE of the U-Net PL. For the ME (image and dose), these tests were used to compare the ME of each method to 0 (null distribution). For the DVHs comparisons across the pCT generation methods, a nonparametric permutation test was performed (62) to control the presence of false positives in case of multiple statistical tests (5 cGy DVH bin-wise). In this case, 1000 permutations were performed where for each permutation *i*, randomly selected DVHs were swapped (CTref <->pCT) and the average difference was computed for each dose-bin. For each permuted sample and the original sample, the average difference was then normalized to the standard deviation computed over all the 1000 permutations and the maximum observed difference was selected

as test-statistic (*TS*). A distribution of *TS* across all the permuted samples ($TS_{i,max}$) was obtained and compared to the one from the observed sample (TS_{max}). The adjusted p-value was therefore computed as the probability of having a TS_{max} greater than the $TS_{i,max}$ compared with a significance level of 5% ($p \leq 0.05$). The corresponding percentile over the distribution of all the $TS_{i,max}$ gives a threshold value which determines the dose DVH bins where statistically significant dose difference arises. Unlike bin-wise tests, permutation test gives a single number that summarizes the discrepancy of the DVHs between the two groups, rather than the discrepancy of a particular bin and, therefore, accounts for multiple comparisons. The mathematical formulation of the permutation test can be found in Chen et al. (63). The test allowed thus to report a robust bin-wise comparison across DVH value of each method, but also to compare the lowest MAE among all the methods to the MAE of each method and the ME of each method to 0.

The Friedman test was used to compare the MAE or the ME of each pCT method between the three different training (1, 2 and 3) (Appendix 5). Results were considered as significant when $p \leq 0.05$.

RESULTS

Imaging endpoints and calculation time

Examples of MRI, CT_{ref}, and pCTs generated by each method are illustrated in Additional Fig. 1.

Table 1 lists the imaging endpoints corresponding to each pCT generation method for the VOIs. The GAN L2 and U-Net L2 showed the lowest MAE and ME (in absolute value) for soft tissue and bone. The GAN PL showed significant lower MAE for the whole pelvis and the soft tissue, than the U-Net PL. The PBM provided the highest corresponding values. Except for the bone, the MEs of GAN L2 and U-Net L2 were not significantly different from a null distribution. Assessing the stochastic effect, the three measurements by method confirmed that GAN L2 and U-Net L2 provided the lowest image uncertainties (Appendix 5).

The mean calculation time to generate one pCT was 15 s for the DLMs and 62 min for the PBM (without using cluster architecture or GPU parallelization).

Dosimetric endpoints

Fig. 2 shows the mean DVHs for the CT_{ref} and each method, by VOI. No DVH points significantly differed when comparing GAN L2 or U-Net L2 DVHs and CT_{ref} DVHs. Most of the points with significant differences were observed for the PBM, GAN PL, and U-Net PL.

Fig. 3 displays the dose uncertainties (MAE) of each method along the DVHs by VOI. GAN L2 provided the lowest dose uncertainties, compared with the other methods. The PBM presented the highest dose uncertainties. Additional Fig. 2 displays the systematic dose uncertainties (ME) of each method along the DVHs, by VOI. The GAN L2 and U-Net L2 presented the lowest ME (in absolute value). The ME of these methods were not significantly different from a null distribution, along the DVH. The PBM, GAN PL, and U-Net PL provided the highest ME (in absolute value). Table 2 lists the mean doses to target volume and OARs and dose uncertainties (MAE) and systematic dose uncertainties (ME) for specific DVH points. The GAN L2 and U-Net L2 showed the lowest MAE and ME. No statistically significant differences were found between MAE of GAN PL and U-Net PL.

Table 3 displays the mean gamma and gamma passrate values calculated from the CT_{ref} and pCT dose distributions for each method. The highest mean gamma values were found for the U-Net L2 and GAN L2. The lowest gamma-pass rate and highest mean gamma values were found for the PBM.

Additional Fig. 3 illustrates the pCTs, dose distributions and gamma maps obtained from a patient.

DISCUSSION

A total of six DLMs for pelvis pCT generation from MRI were investigated and compared with a PBM. Several hyperparameters of the DLMs were optimized according to imaging endpoints (Appendix 4). Compared to the CT_{ref}, the pCTs generated by DLMs and PBM provided overall low dose uncertainties, thereby making them clinically acceptable for MRI-based prostate dose planning (Fig. 2). Regarding dose accuracy and calculation time, in comparison with PBM,

DLMs appear particularly promising for clinical use. Among DLMs, the most accurate methods are GAN L2 and U-Net L2 (Table 2, Fig. 2, Fig. 3 and Additional Fig. 2).

Deep learning has been used for pCT generation from MRI exclusively in the brain (23, 25, 26, 57) and pelvis (27, 29, 48, 64–66). In the pelvis, four deep learning architectures have used: fully convolutional network (FCN) (65), deep embedding convolutional neural network (DECNN) (66), U-Net (48, 64), and GAN architecture without perceptual loss (27, 29). Imaging and dose endpoints have been considered to evaluate these methods within the scope of radiotherapy. All six studies evaluated the imaging endpoints in cohorts ranging from 15 to 39 patients, among which this of Arabi et al. (48) used the same cohort of patient than used in the present study. In the entire pelvis, the MAEs were 42.4 HU (65) and 42.5 HU (66) when FCN and DECNN architectures were used, respectively. Using a U-Net architecture, the MAEs were 30 HU (64) and 32.7 HU (48). Using a GAN architecture, the MAEs were 60 HU (27) and 39.0 HU (29). Although, the comparison can only be indirect, our proposed GAN L2 and U-Net L2 DLM compared favorably with a MAE value of 34.1 HU and 34.4 HU, respectively.
Only four studies in the literature evaluated the dose uncertainties, three in the pelvis (27, 48, 64) and one in the brain (26), considering various dosimetric endpoints. In the pelvis, the mean dose uncertainties reported using U-Net and GAN DLMs were lower than 0.2% and 0.5% in all the VOIs (27, 48, 64). Our mean dose uncertainties with GAN DLMs appears comparable (Table 2). In the literature, the reported mean gamma pass-rates were 98% (64) and 95% (48) with a 1%/1 mm criteria (in 2D), and 95% with a 2%/2 mm criteria (in 3D) (27). In comparison, we obtained a higher gamma pass-rate (99%) with our GAN and U-Net DLMs (Table 3).

In our study, compared to the PBM, our GAN and U-Net DLMs provided lower imaging uncertainties, with the lowest for GAN L2 and U-Net L2 (Table 1). The perceptual loss in U-Net and GAN did not decrease the HU uncertainty. This may be explained by the choice of our evaluation metric (HU difference, required within a dose calculation perspective), and not considering the image quality metrics (Universal image Quality Index (UQI) (67)) such as Peak Signal-to-Noise Ratio (PSNR), Normalized Mutual Information (NMI) (68), Structural SIMilarity (SSIM) (69), Visual Information Fidelity (VIF) (70) and Learned Perceptual Image Patch Similarity (LPIPS) (47, 71) used in computer vision applications. We did not consider any other image quality metrics than MAE and ME in this study, since not impacting dose calculation.

While the perceptual loss does not seem to provide any advantage for dose calculation, this loss function may be relevant for other image processing task like segmentation and registration within a CBCT-based IGRT. Moreover, for the bone, the addition of adversarial term tends to decrease the imaging uncertainty in the GAN.
Considering all the methods, the largest uncertainties were observed for the bone (up to 144 HU for MAE), which are related to the highest HU values in the bone (345 HU, additional Table 1). For the rectum, large uncertainties were also observed (up to 78 HU for MAE, Table 1), potentially related to the difference in gas pocket between the MRI and CT_{ref}. However, all these methods seemed to incorrectly reproduce the real air pockets (when they were present both on CT and MRI), as illustrated in Additional Fig. 1 (sagittal views). This issue could be explained by the complex detection of air pocket with the T2 MRI and lack of variability of air pockets in the training cohort.

GAN PL and GAN L2 provided significant lower imaging uncertainties (MAE) than U-Net PL and U-Net L2, respectively. GAN L2 and U-Net L2 presented the lowest dose uncertainties (MAE) (Table 2 and Fig. 3) without any systematic dose uncertainties (ME) (Table 2 and additional Fig. 2). Nevertheless, these results appeared more robust with the adversarial term of the GAN discriminator loss function (Appendix 5). In our previous study that compared the bulk density method (BDM), atlas-based method (ABM) and PBM, PBM was found to be the most accurate pCT generation method. Additional Fig. 4 compares the nine strategies in the whole series of patients (BDM, ABM, PBM, and the six DLMs). This figure confirms that GAN L2 and U-Net L2 are the most accurate methods and ABM and BDM are the least accurate. Overall, the dose uncertainties of the pCTs of each method are small, unlikely to be clinically relevant in terms of local control and toxicity.

Our study presents some limitations. First, before the learning process, non-rigid registration was used to align pelvic anatomies between MRI and CT_{ref}, with the intrinsic uncertainties linked to the deformable image registration algorithm. However, we previously quantified these geometrical uncertainties in (20) by calculating the Dice scores before registration ($CT_{initial}$ vs MRI) and after registration (CT_{ref} vs MRI) for the prostate, seminal vesicles, bladder, and rectum volumes. We found that all Dice scores were significantly improved by the non-rigid registration ($p \leq 0.05$). Furthermore, these registrations did not correct the gas volatility in the

digestive structures. The dose uncertainties related to rectal variations were quantified in our previous study using the PBM (20). The gas correction (gas inside the pCT was deleted and replaced by the gas from the CT_{ref}) yielded a significant lowest dose uncertainty for the rectum between V15 Gy and V25 Gy. Second, we investigated only the T_2-weighted MRI sequences. DLMs may be sensitive to variations in MRI and other MR sequences could have been used. Because of the relative low number of patients, the optimization was performed with only one of the three draws. No test set was therefore used, potentially exposing our optimization to a bias. Even if the pCTs generated by DLMs and PBM provided overall low dose uncertainties, an outlier analysis should be however performed on an independent and large enough dataset. The GAN DLM was trained with 2D axial slices and not in with 3D images due to the GPU memory limitations. To overcome this issue, 3D patches could have been used during the training however at the expense of the contextual information inclusion. Indeed, small 3D patches (32x32x32 or 64x64x64) ignore the global anatomical information, as opposed to a 2D slice. In addition, 3D architectures are often shallow compared to 2D architectures (29). Another solution could be brought by the generation of pCTs from individual 2D axial, sagittal, and coronal slices fused together, adding 30 more seconds once the networks are trained. Finally, other emerging deep learning architectures such as the cycle-GAN, which may have allowed to overcome some intra-individual co-registration issues, could have been investigated.

CONCLUSION

To generate pCT for MRI-based prostate dose planning, deep learning methods appear to be particularly promising for clinical practice owing to the low dose uncertainty and fast calculation time. The U-Net and GAN DLMs with L2 loss function provide the lowest dose uncertainties. These MRI approaches in prostate cancer radiotherapy, which do not require any CT, could thereby improve the accuracy of VOI delineation and can also be used for (re)planning in the MRI-LINAC workflow (72).

REFERENCES

1. Pathmanathan AU, McNair HA, Schmidt MA, *et al.* Comparison of prostate delineation on multimodality imaging for MR-guided radiotherapy. *Br. J. Radiol.* 2019;92:20180948.

2. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat. Oncol.* 2017;12.

3. Johnstone E, Wyatt JJ, Henry AM, *et al.* Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy. *Int. J. Radiat. Oncol.* 2018;100:199–217.

4. Lambert J, Greer PB, Menk F, *et al.* MRI-guided prostate radiation therapy planning: Investigation of dosimetric accuracy of MRI-based dose planning. *Radiother. Oncol.* 2011;98:330–334.

5. Lee YK, Bollet M, Charles-Edwards G, *et al.* Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone. *Radiother. Oncol.* 2003;66:203–216.

6. Hoogcarspel SJ, Van der Velden JM, Lagendijk JJ, *et al.* The feasibility of utilizing pseudo CT-data for online MRI based treatment plan adaptation for a stereotactic radiotherapy treatment of spinal bone metastases. *Phys. Med. Biol.* 2014;59:7383.

7. Chen L, Nguyen T-B, Jones É, *et al.* Magnetic Resonance–Based Treatment Planning for Prostate Intensity-Modulated Radiotherapy: Creation of Digitally Reconstructed Radiographs. *Int. J. Radiat. Oncol. • Biol. • Phys.* 2007;68:903–911.

8. Chin AL, Lin A, Anamalayil S, *et al.* Feasibility and limitations of bulk density assignment in MRI for head and neck IMRT treatment planning. *J. Appl. Clin. Med. Phys.* 2014;15.

9. Gudur MSR, Hara W, Le Q-T, *et al.* A unifying probabilistic Bayesian approach to derive electron density from MRI for radiation therapy treatment planning. *Phys. Med. Biol.* 2014;59:6595–6606.

10. Dowling JA, Sun J, Pichler P, *et al.* Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences. *Int. J. Radiat. Oncol. Biol. Phys.* 2015;93:1144–1153.

11. Burgos N, Cardoso MJ, Guerreiro F, *et al.* Robust CT Synthesis for Radiotherapy Planning: Application to the Head and Neck Region. In: Navab N, Hornegger J, Wells WM, *et al.*, eds. *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015*. Lecture Notes in Computer Science. Springer International Publishing; 2015:476–484.

12. Dowling JA, Lambert J, Parker J, *et al.* An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 2012;83:e5–e11.

13. Guerreiro F, Burgos N, Dunlop A, *et al.* Evaluation of a multi-atlas CT synthesis approach for MRI-only radiotherapy treatment planning. *Phys. Med.* 2017;35:7–17.

14. Sjölund J, Forsberg D, Andersson M, *et al.* Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Phys. Med. Biol.* 2015;60:825–839.

15. Uh J, Merchant TE, Li Y, *et al.* MRI-based treatment planning with pseudo CT generated through atlas registration. *Med. Phys.* 2014;41:051711.

16. Arabi H, Koutsouvelis N, Rouzaud M, *et al.* Atlas-guided generation of pseudo-CT images for MRI-only and hybrid PET–MRI-guided radiotherapy treatment planning. *Phys. Med. Biol.* 2016;61:6531–6552.

17. Persson E, Gustafsson C, Nordström F, *et al.* MR-OPERA: A Multicenter/Multivendor Validation of Magnetic Resonance Imaging–Only Prostate Treatment Planning Using Synthetic Computed Tomography Images. *Int. J. Radiat. Oncol.* 2017;99:692–700.

18. Andreasen D, Van Leemput K, Edmund JM. A patch-based pseudo-CT approach for MRI-only radiotherapy in the pelvis. *Med. Phys.* 2016;43:4742.

19. Aouadi S, Vasic A, Paloor S, *et al.* Generation of synthetic CT using multi-scale and dual-contrast patches for brain MRI-only external beam radiotherapy. *Phys. Med.* 2017;42:174–184.

20. Largent A, Barateau A, Nunes J-C, *et al.* Pseudo-CT generation for MRI-only radiotherapy treatment planning: comparison between patch-based, atlas-based, and bulk density methods. *Int. J. Radiat. Oncol.* 2018.

21. Andreasen D, Van Leemput K, Hansen RH, *et al.* Patch-based generation of a pseudo CT from conventional MRI sequences for MRI-only radiotherapy of the brain. *Med. Phys.* 2015;42:1596–1605.

22. Shafai-Erfani G, Wang T, Lei Y, *et al.* Dose evaluation of MRI-based synthetic CT generated using a machine learning method for prostate cancer radiotherapy. *Med. Dosim.* 2019.

23. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med. Phys.* 2017;44:1408–1419.

24. Nie D, Trullo R, Lian J, *et al.* Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In: *Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017*. Lecture Notes in Computer Science. Springer, Cham; 2017:417–425.

25. Wolterink JM, Dinkla AM, Savenije MHF, *et al.* Deep MR to CT Synthesis Using Unpaired Data. In: *Simulation and Synthesis in Medical Imaging*. Lecture Notes in Computer Science. Springer, Cham; 2017:14–23.

26. Dinkla AM, Wolterink JM, Maspero M, *et al.* MR-only brain radiotherapy: Dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network. *Int. J. Radiat. Oncol. • Biol. • Phys.* 2018;0.

27. Maspero M, Savenije MHF, Dinkla AM, *et al.* Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys. Med. Biol.* 2018;63:185001.

28. Fu J, Yang Y, Singhrao K, *et al.* Male pelvic synthetic CT generation from T1-weighted MRI using 2D and 3D convolutional neural networks. *ArXiv180300131 Phys.* 2018.

29. Nie D, Trullo R, Lian J, *et al.* Medical Image Synthesis with Deep Convolutional Adversarial Networks. *IEEE Trans. Biomed. Eng.* 2018;65:2720–2730.

30. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.

31. Meyer P, Noblet V, Mazzara C, *et al.* Survey on deep learning for radiotherapy. *Comput. Biol. Med.* 2018;98:126–146.

32. Higaki T, Nakamura Y, Tatsugami F, *et al.* Improvement of image quality at CT and MRI using deep learning. *Jpn. J. Radiol.* 2018.

33. Alkadi R, Taher F, El-baz A, *et al.* A Deep Learning-Based Approach for the Detection and Localization of Prostate Cancer in T2 Magnetic Resonance Images. *J. Digit. Imaging*. 2018.

34. Laukamp KR, Thiele F, Shakirin G, *et al.* Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. *Eur. Radiol.* 2019;29:124–132.

35. Liang S, Tang F, Huang X, *et al.* Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur. Radiol.* 2018.

36. Nyflot MJ, Thammasorn P, Wootton LS, *et al.* Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. *Med. Phys.* 2018.

37. Sahiner B, Pezeshk A, Hadjiiski LM, *et al.* Deep learning in medical imaging and radiation therapy. *Med. Phys.* 2019;46:e1–e36.

38. Gong K, Yang J, Kim K, *et al.* Attenuation correction for brain PET imaging using deep neural network based on Dixon and ZTE MR images. *Phys. Med. Biol.* 2018;63:125011.

39. Ladefoged CN, Marner L, Hindsholm A, *et al.* Deep Learning Based Attenuation Correction of PET/MRI in Pediatric Brain Tumor Patients: Evaluation in a Clinical Setting. *Front. Neurosci.* 2019;12.

40. Leynes AP, Yang J, Wiesinger F, *et al.* Zero-Echo-Time and Dixon Deep Pseudo-CT (ZeDD CT): Direct Generation of Pseudo-CT Images for Pelvic PET/MRI Attenuation Correction Using Deep Convolutional Neural Networks with Multiparametric MRI. *J. Nucl. Med.* 2018;59:852–858.

41. Torrado-Carvajal A, Vera-Olmos J, Izquierdo-Garcia D, *et al.* Dixon-VIBE Deep Learning (DIVIDE) Pseudo-CT Synthesis for Pelvis PET/MR Attenuation Correction. *J. Nucl. Med.* 2019;60:429–435.

42. Kläser K, Markiewicz P, Ranzini M, *et al.* Deep Boosted Regression for MR to CT Synthesis. In: Gooya A, Goksel O, Oguz I, *et al.*, eds. *Simulation and Synthesis in Medical Imaging*.Vol 11037. Cham: Springer International Publishing; 2018:61–70.

43. Nyholm T, Svensson S, Andersson S, *et al.* MR and CT data with multiobserver delineations of organs in the pelvic area-Part of the Gold Atlas project. *Med. Phys.* 2018;45:1295–1300.

44. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al.* Generative Adversarial Networks. *ArXiv14062661 Cs Stat*. 2014.

45. Yang Q, Yan P, Zhang Y, *et al.* Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. *IEEE Trans. Med. Imaging*. 2018;37:1348–1357.

46. Wang C, Xu C, Wang C, *et al.* Perceptual Adversarial Networks for Image-to-Image Transformation. *IEEE Trans. Image Process.* 2018;27:4066–4079.

47. Armanious K, Jiang C, Fischer M, *et al.* MedGAN: Medical Image Translation using GANs. *ArXiv180606397 Cs*. 2018.

48. Arabi H, Dowling JA, Burgos N, *et al.* Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region. *Med. Phys.* 2018;45:5218–5233.

49. Rivest-Hénault D, Dowson N, Greer PB, *et al.* Robust inverse-consistent affine CT–MR registration in MRI-assisted and MRI-alone prostate radiation therapy. *Med. Image Anal.* 2015;23:56–69.

50. Rivest-Hénault D, Greer P, Fripp J, *et al.* Structure-Guided Nonrigid Registration of CT–MR Pelvis Scans with Large Deformations in MR-Based Image Guided Radiation Therapy. In: Erdt M, Linguraru MG, Laura CO, *et al.*, eds. *Clinical Image-Based Procedures. Translational Research in Medical Imaging*. Lecture Notes in Computer Science. Springer International Publishing; 2013:65–73.

51. Wachinger C, Brennan M, Sharp GC, *et al.* Efficient Descriptor-Based Segmentation of Parotid Glands With Nonlocal Means. *IEEE Trans. Biomed. Eng.* 2017;64:1492–1502.

52. Silpa-Anan C, Hartley R. Optimised KD-trees for fast image descriptor matching. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition.*; 2008:1–8.

53. Ibanez L, Schroeder W, Ng L, *et al.* The ITK software guide. 2005.

54. Hahnloser RHR, Sarpeshkar R, Mahowald MA, *et al.* Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*. 2000;405:947–951.

55. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs*. 2014.

56. Mao X, Li Q, Xie H, *et al.* Least Squares Generative Adversarial Networks. :9.

57. Emami H, Dong M, Nejad-Davarani SP, *et al.* Generating Synthetic CTs from Magnetic Resonance Images using Generative Adversarial Networks. *Med. Phys.* 0.

58. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*. 2014.

59. Anon. *Deep Learning for humans. Contribute to keras-team/keras development by creating an account on GitHub*. Keras; 2018.

60. Beckendorf V, Guerif S, Prisé EL, *et al.* 70 Gy Versus 80 Gy in Localized Prostate Cancer: 5-Year Results of GETUG 06 Randomized Trial. *Int. J. Radiat. Oncol. • Biol. • Phys.* 2011;80:1056–1063.

61. Marks LB, Yorke ED, Jackson A, *et al.* Use of normal tissue complication probability models in the clinic. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;76:S10–S19.

62. Ross SM. Nonparametric Hypotheses Tests. In: ; 2010.

63. Chen C, Witte M, Heemsbergen W, *et al.* Multiple comparisons permutation test for image based data mining in radiotherapy. *Radiat. Oncol.* 2013;8.

64. Chen S, Qin A, Zhou D, *et al.* Technical Note: U-net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning. *Med. Phys.* 2018;45:5659–5665.

65. Nie D, Cao X, Gao Y, *et al.* Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks. In: *Deep Learning and Data Labeling for Medical Applications*. Lecture Notes in Computer Science. Springer, Cham; 2016:170–178.

66. Xiang L, Wang Q, Nie D, *et al.* Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. *Med. Image Anal.* 2018;47:31–44.

67. Zhou Wang, Bovik AC. A universal image quality index. *IEEE Signal Process. Lett.* 2002;9:81–84.

68. Studholme C, Hill DLG, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit.* 1999;32:71–86.

69. Wang Z, Bovik AC, Sheikh HR, *et al.* Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 2004;13:600–612.

70. Sheikh HR, Bovik AC. IMAGE INFORMATION AND VISUAL QUALITY. :4.

71. Zhang R, Isola P, Efros AA, *et al.* The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE; 2018:586–595.

72. Bird D, Henry AM, Sebag-Montefiore D, *et al.* A Systematic Review of the Clinical Implementation of Pelvic Magnetic Resonance Imaging (MR)-Only Planning for External Beam Radiation Therapy. *Int. J. Radiat. Oncol.* 2019.

Tables and figures

Table 1. Imaging endpoints comparing the reference CT to the pseudo-CTs obtained by each method for the entire pelvis, soft tissue, and bone

MAE: mean absolute error of HU values defined as the mean difference (in absolute value) of HU values per voxel between the reference CT and the pseudo-CT and; ME: mean error, defined as the mean difference of HU values per voxel between the reference CT and the pseudo-CT of each method.

The imaging endpoint values are expressed as mean ± standard deviation. The Wilcoxon test was used to: firstly, compare the MAE of the GAN with L2 loss to those of the other methods; and to secondly, compare the ME of the methods to a null distribution. Significant differences ($p \leq 0.05$) are displayed using the symbol *. The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.

Table 2. Reference dose values, dose uncertainties (MAE) and systematic dose uncertainties (ME) for each pseudo-CT generation method for each volume of interest

The mean values of DVH points are reported for the reference CT. The dose uncertainty is defined as the mean absolute DVH differences between the DVH calculated from the reference CT and those obtained from the pCTs. The systematic dose uncertainty is defined as the mean DVH differences between the DVH calculated from the reference CT and those obtained from the pCTs. The Wilcoxon test was used to: firstly, compare the dose uncertainty (MAE) of the GAN with L2 loss to those of the other methods; and secondly, to compare the systematic dose uncertainty (ME) of the methods to a null distribution. Significant differences ($p \leq 0.05$) are displayed using the symbol *. The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.

Table 3. Mean gamma and gamma pass-rate calculated from the reference CT and pseudo-CT dose distributions according to each method

Values are mean ± standard deviation.

The Wilcoxon test was used to compare the gamma values of the GAN with L2 loss to those of the other methods. Significant differences ($p \leq 0.05$) are displayed using the symbol *.

The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.

Fig. 1. U-Net and GAN deep learning trained architectures with different implemented loss functions

"I" corresponds to the training MRI and "C" to the corresponding training CT.

Two deep learning neural networks (U-Net and GAN) were trained with four loss functions (L2 loss, single-scale perceptual loss, multi-scale perceptual (MP) loss and weighted multi-scale perceptual (WMP) loss) yielding six different deep learning training strategies: U-Net with L2 loss (U-Net L2), U-Net with single-scale perceptual loss (U-Net PL), GAN with L2 loss (GAN L2), GAN with single-scale perceptual loss (GAN PL), GAN with multi-scale perceptual loss (GAN MPL), and GAN with weighted multi-scale perceptual loss (GAN WMPL)**.** For each patient from the training database, the CT and MRI training images were first non-rigidly co-registered. The DLM architecture of the U-Net was symmetric, with N encoding and decoding units each. The contracting path consisted of 12 3×3 convolution layers with stride 2 for down-sampling, each followed by batch normalization and ReLU activation function. To train the U-Net DLM two different loss functions were implemented: L2 loss and single-scale perceptual loss. The VGG16 network was used to compute the features inside the CT and pCT images.

The training of the GAN consists of two competing multilayer networks: the generator and the discriminator. The generator is used as a regression model to provide pCTs from MRIs. The generator employed in this study has the same architecture than the previous described U-Net. The discriminator aims to distinguish the real image (ground truth) from the realistic fake image (pCT) produced by the generator. The GANs are formulated mathematically as a minimax game between these two networks, which is solved by alternating gradient optimization. The input data of the generator are MRI and registered CT images that provide pCTs. Then, the discriminator classifies these pCTs as real or fake CTs until the discriminator cannot determine

whether the pCT looks like a real CT or not. In the testing step, for a new given test patient, the MRI goes through the trained network to obtain the corresponding pCT.

Fig. 2. Mean DVHs for prostate PTV, bladder, rectum, and femoral heads from the reference CT and pseudo-CTs generated by each method

PBM: patch-based method; U-Net L2: U-Net using a L2 loss method; U-Net PL: U-Net using a single-scale perceptual loss (layer 7) method; GAN L2: Generative Adversarial Network using a L2 loss method; GAN PL: Generative Adversarial Network using a single-scale perceptual loss (layer 7) method; GAN MPL: Generative Adversarial Network using an multi-scale perceptual loss method; GAN WMPL: Generative Adversarial Network using a weighted multi-scale perceptual loss method;

Permutation tests were used to compare the DVHs from the reference CT to those of the pseudo-CT generation methods. Significant differences ($p \leq 0.05$) between the DVHs are displayed at the top of each figure using the symbol *.

Fig. 3. Dose uncertainties (MAE) for all pseudo-CT generation methods along the entire DVH for the prostate PTV, bladder, rectum and femoral heads

PBM: patch-based method; U-Net L2: U-Net using a L2 loss method; U-Net PL: U-Net using a single-scale perceptual loss (layer 7) method; GAN L2: Generative Adversarial Network using a L2 loss method; GAN PL: Generative Adversarial Network using a single-scale perceptual loss (layer 7) method; GAN MPL: Generative Adversarial Network using an multi-scale perceptual loss method; GAN WMPL: Generative Adversarial Network using a weighted multi-scale perceptual loss method;

The dose uncertainty is defined as the mean absolute DVH differences between the reference CT and the pCT corresponding to each method. Permutation tests were used to compare the absolute DVH differences of the GAN L2 method to those of the other methods. Significant differences ($p \leq 0.05$) are displayed at the top of each figure with *.

Training data

MRI

CT

Co-registered MRI-CT

I

C

Training networks

C

C

Loss functions

1 L2 loss

2 Perceptual loss

I

Encoding/decoding networks (U-Net)

U-Net(I)

pseudo-CT

VGG(U-Net(I))
VGG(C)

U-Net Network

C

Loss functions

3 L2 loss

4 Perceptual loss

5 MP loss

6 WMP loss

I

Generator (G)

G(I)

pseudo-CT

VGG(G(I))
VGG(C)

Discriminator (D)

Real CT (1)

Fake CT (0)

D(I)

GAN Network

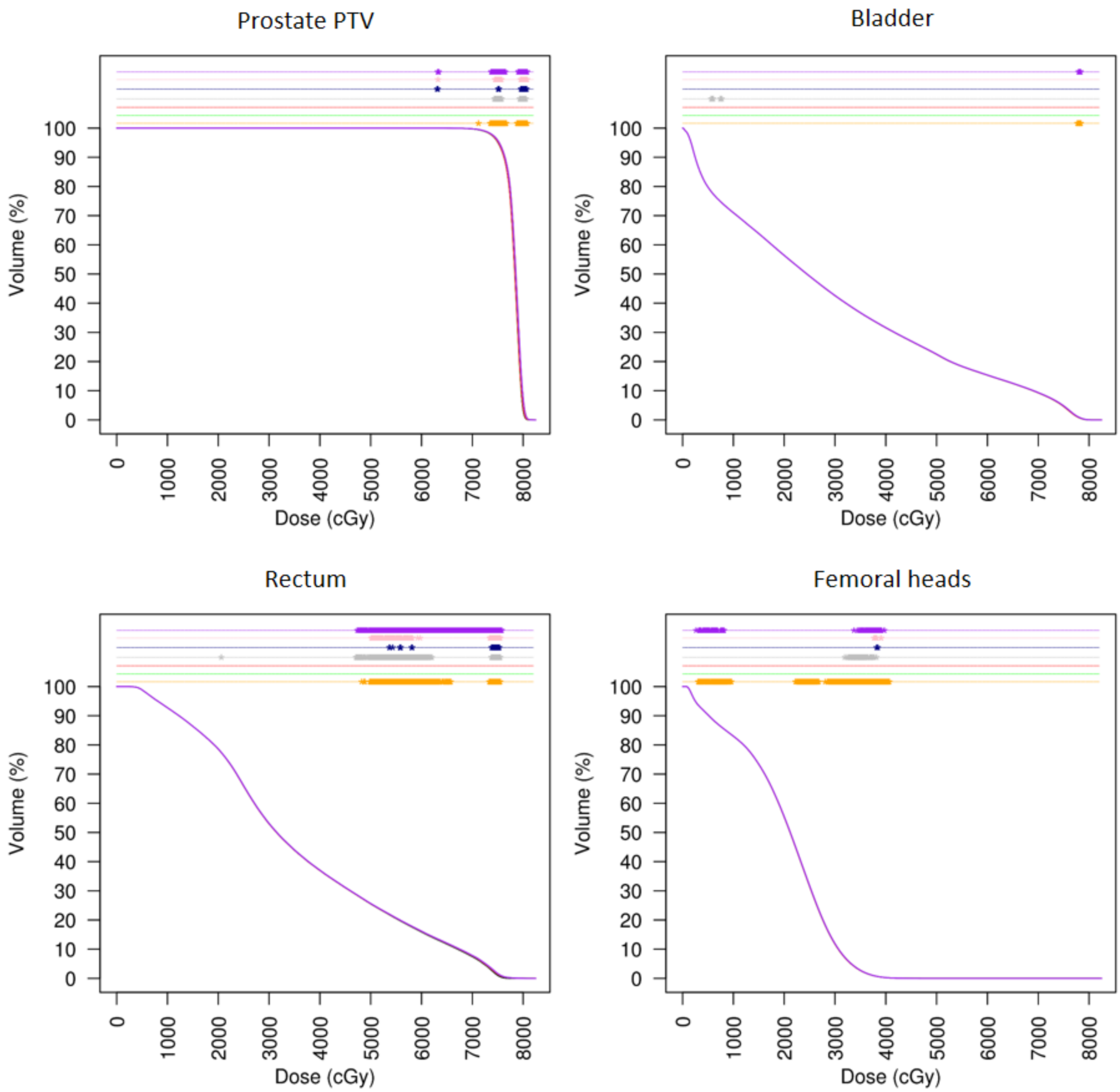
Testing data

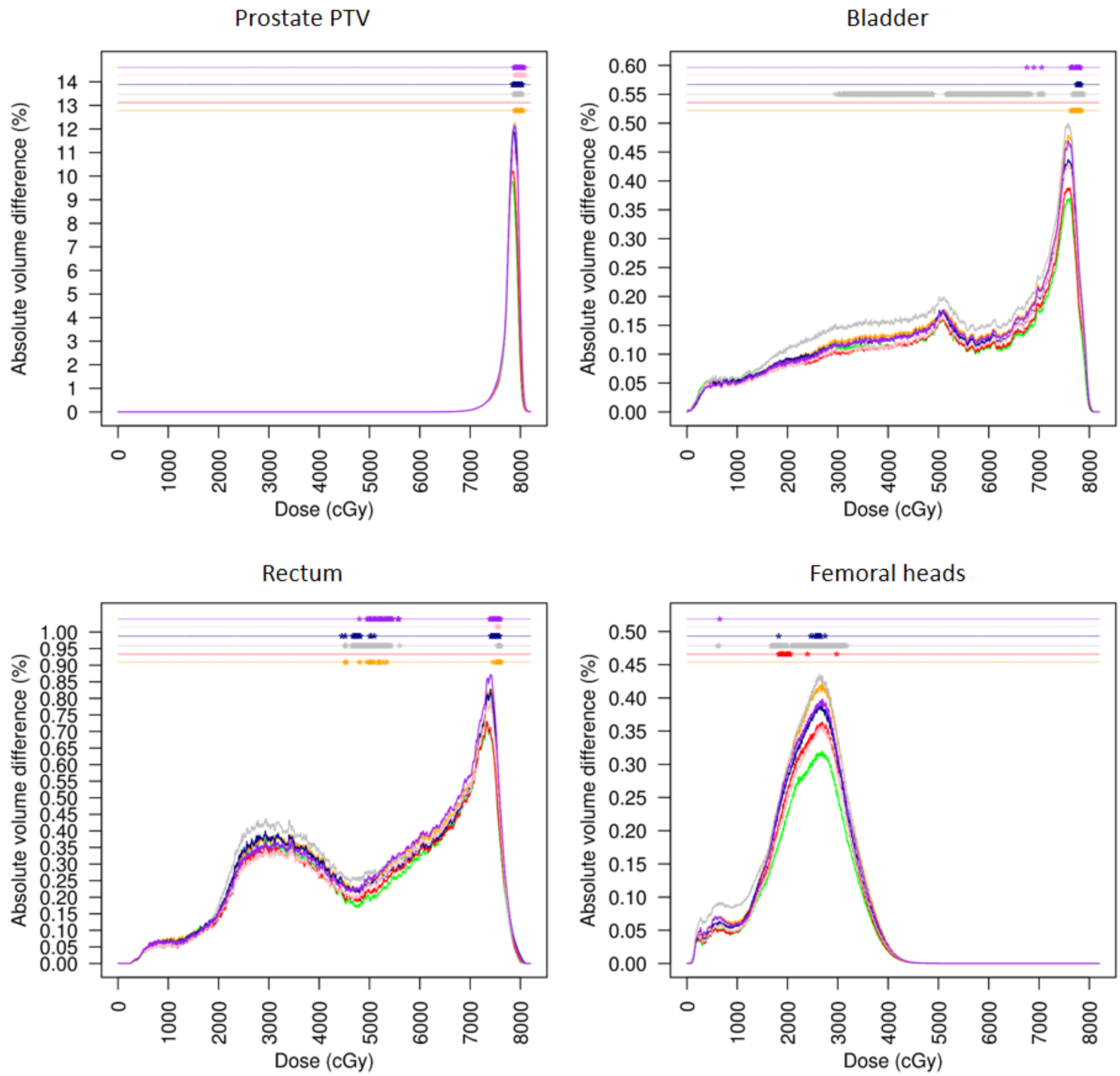
MRI

Trained Network

U-Net 1-2 or **GAN** 3-6

pseudo-CT


Prostate PTV
Bladder
Rectum
Femoral heads
Volume (%)
Dose (cGy)
0
10
20
30
40
50
60
70
80
90
100
0
1000
2000
3000
4000
5000
6000
7000
8000
CT
Patch-based method
U-Net L2 method
U-Net PL method
GAN L2 method
GAN PL method
GAN MPL method
GAN WMPL method
Wilcoxon: CT vs GAN PL method
Wilcoxon: CT vs GAN L2 method
Wilcoxon: CT vs U-Net L2 method
Wilcoxon: CT vs patch-based method
Wilcoxon: CT vs GAN WMPL method
Wilcoxon: CT vs GAN MPL method
Wilcoxon: CT vs U-Net PL method


Prostate PTV
Absolute volume difference (%)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0
1000
2000
3000
4000
5000
6000
7000
8000
Dose (cGy)
Bladder
Absolute volume difference (%)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Dose (cGy)
Rectum
Absolute volume difference (%)
Dose (cGy)
Femoral heads
Absolute volume difference (%)
Dose (cGy)
Patch-based method
U-Net L2 method
U-Net PL method
GAN L2 method
GAN PL method
GAN MPL method
GAN WMPL method
Wilcoxon: GAN L2 method vs patch-based method
Wilcoxon: GAN L2 method vs U-Net L2 method
Wilcoxon: GAN L2 method vs U-Net PL method
Wilcoxon: GAN L2 method vs GAN PL method
Wilcoxon: GAN L2 method vs GAN MPL method
Wilcoxon: GAN L2 method vs GAN WMPL method

		Endpoints (HU)	Methods used to generate pseudo-CT						
			Patch-based method	U-Net methods		GAN methods			
				L2 loss	Single-scale perceptual loss (layer 7)	L2 loss	Single-scale perceptual loss (layer 7)	Multi-scale perceptual loss	Weighted multi-scale perceptual loss
Entire pelvis		MAE	44.7 ± 11.4*	34.4 ± 7.7	36.8 ± 6.0* ◦	34.1 ± 7.5	34.9 ± 6.4*	35.6 ± 6.2*	35.1 ± 6.8*
		ME	9.9 ± 18.1*	-1.0 ± 14.2	3.3 ± 13.6	-1.1 ± 13.7	4.1 ± 13.9	1.9 ± 13.3	1.2 ± 14.0
Soft tissue only	Entire soft tissues	MAE	36.4 ± 11.3*	26.7 ± 6.4	29.2 ± 5.2* ◦	26.5 ± 6.4	27.1 ± 5.3*	27.8 ± 5.0*	27.4 ± 5.6*
		ME	6.0 ± 19.0	-2.6 ± 14.7	0.9 ± 14.0	-2.8 ± 14.3	1.3 ± 14.8	-0.6 ± 14.1	-1.2 ± 14.8
	Prostate (CTV)	MAE	20.6 ± 6.0*	18.1 ± 5.2	22.2 ± 4.9* ◦	17.7 ± 4.49	23.3 ± 5.9*	21.6 ± 3.7*	22.9 ± 5.8*
		ME	8.2 ± 15.0*	0.8 ± 12.9	14.4 ± 11.5*	0.3 ± 12.0	16.8 ± 11.5*	12.3 ± 11.2*	13.9 ± 13.8*
	Bladder	MAE	21.1 ± 9.0*	18.6 ± 7.4	19.3 ± 10.0	18.8 ± 8.9	19.6 ± 9.3	20.2 ± 10.0	19.9 ± 9.3*
		ME	10.7 ± 14.0*	3.4 ± 13.6	5.3 ± 16.6*	3.7 ± 14.6	7.7 ± 15.5*	3.4 ± 16.4	5.7 ± 16.4*
	Rectum	MAE	78.0 ± 60.5*	65.0 ± 65.7	68.6 ± 66.1°	68.3 ± 64.4	72.9 ± 68.6	69.2 ± 65.5	71.3 ± 68.5
		ME	7.0 ± 73.2*	-24.0 ± 72.5	-17.5 ± 74.1	-20.5 ± 73.6	-11.3 ± 78.9	-16.6 ± 76.3	-16.0 ± 77.2
Bone only	Whole pelvic bone	MAE	143.6 ± 27.8*	125.3	126.3 ± 22.1*	123.9 ± 20.6	127.9 ± 22.3*	127.1 ± 21.1*	126.7 ± 21.2*

				± 22.0 *					
		ME	58.3 ± 45.5*	20.2 ± 42.3 *	32.7 ± 41.8*	19.4 ± 41.4 *	39.7 ± 40.8*	31.8 ± 41.4*	28.8 ± 41.3*
	Femoral heads	MAE	109.3 ± 27.0*	102.0 ± 24.4 *	103.8 ± 22.5*	100.2 ± 20.4	104.7 ± 21.5*	104.9 ± 19.2*	104.6 ± 20.9*
		ME	36.5 ± 49.9*	5.0 ± 49.5	21.9 ± 48.8*	5.1 ± 47.2	29.8 ± 48.0*	16.9 ± 48.1*	19.6 ± 48.0*

Table 1. Imaging endpoints comparing the reference CT to the pseudo-CTs obtained by each method for the entire pelvis, soft tissue, and bone

MAE: mean absolute error of HU values defined as the mean difference (in absolute value) of HU values per voxel between the reference CT and the pseudo-CT and; ME: mean error, defined as the mean difference of HU values per voxel between the reference CT and the pseudo-CT of each method.

The imaging endpoint values are expressed as mean ± standard deviation. The Wilcoxon test was used to: firstly, compare the MAE of the GAN with L2 loss to those of the other methods; and to secondly, compare the ME of the methods to a null distribution. Significant differences ($p \leq 0.05$) are displayed using the symbol *. The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.

Volumes of interest			**Prostate CTV**		**Prostate PTV**		**Rectum**			**Bladder**			**Femoral heads**	
Dosimetric endpoints			$D_{99\%}$ (cGy)	D_{mean} (cGy)	$V_{95\%}$ (%)	D_{mean} (cGy)	V_{70Gy} (%)	D_{max} (cGy)	D_{mean} (cGy)	V_{50Gy} (%)	D_{max} (cGy)	D_{mean} (cGy)	V_{30Gy} (%)	D_{mean} (cGy)
Reference CT values			7628 ± 50	7869 ± 52	97.1 ± 1.4	7816 ± 47	7.5 ± 3.0	7331 ± 166	3603 ± 277	22.4 ± 11.5	7784 ± 101	2951 ± 981	11.8 ± 6.5	1992 ± 249
Dose uncertainties (MAE)	**Patch-based method**		39 ± 24*	36 ± 22*	0.6 ± 0.6	35 ± 20*	0.5 ± 0.9	48 ± 58	19 ± 15*	0.2 ± 0.1	32 ± 22*	12 ± 10*	0.3 ± 0.3*	7 ± 6*
	U-Net deep learning method	**L2 loss**	31 ± 31*	29 ± 24*	0.6 ± 0.5	28 ± 23	0.5 ± 0.5	45 ± 45	15 ± 14	0.1 ± 0.1	26 ± 26	9 ± 10	0.3 ± 0.3*	6 ± 5*
		Single-scale perceptual loss (layer 7)	38 ± 23*	35 ± 19*	0.6 ± 0.6	35 ± 17*	0.6 ± 0.8	53 ± 65*	18 ± 14*	0.2 ± 0.1	30 ± 19	10 ± 9	0.3 ± 0.2	7 ± 5*
	GAN deep learning method	**L2 loss**	28 ± 26	26 ± 24	0.6 ± 0.5	26 ± 22	0.5 ± 0.8	45 ± 59	15 ± 13	0.1 ± 0.1	25 ± 23	9 ± 9	0.2 ± 0.2	5 ± 5
		Single-scale perceptual loss (layer 7)	38 ± 24*	36 ± 21*	0.6 ± 0.6	35 ± 19*	0.5 ± 0.8	50 ± 65	18 ± 15*	0.2 ± 0.1	31 ± 19	11 ± 9*	0.3 ± 0.2	7 ± 5
		Multi-scale perceptual loss	34 ± 22*	32 ± 19	0.6 ± 0.6	32 ± 17	0.5 ± 0.8	48 ± 63	16 ± 13	0.1 ± 0.1	28 ± 19	10 ± 8	0.3 ± 0.2	6 ± 5
		Weighted multi-scale perceptual loss	36 ± 22*	34 ± 20*	0.6 ± 0.5	33 ± 18*	0.5 ± 0.8	50 ± 64	18 ± 14*	0.2 ± 0.1	29 ± 19	10 ± 9	0.3 ± 0.2	7 ± 5*

Systematic dose uncertainty (ME)	**Patch-based method**		-16 ± 43*	-12 ± 41	-0.3 ± 0.8*	-13 ± 39*	-0.3 ± 0.8*	-31 ± 69*	-12* ± 21*	-0.1 ± 0.2	-11 ± 37	-5 ± 15*	-0.1 ± 0.4	-2 ± 9*
	U-Net deep learning method	**L2 loss**	1 ± 40	5 ± 38	-0.1 ± 0.8	3 ± 36	-0.2 ± 0.9	-17 ± 73.3	-3 ± 21	0.0 ± 0.2	5.4 ± 36	1 ± 13	0.0 ± 0.3	1 ± 8
		Single-scale perceptual loss (layer 7)	-20 ± 40*	-15 ± 37*	-0.3 ± 0.8*	-16 ± 36*	-0.4 ± 0.9*	-36 ± 76*	-11 ± 21*	0.0 ± 0.2*	-9.6 ± 34	-3* ± 13*	-0.1 ± 0.3	-3* ± 8*
	GAN deep learning method	**L2 loss**	1 ± 38	6 ± 35	-0.1 ± 0.8	4 ± 34	-0.2 ± 0.9	-16 ± 73	-3 ± 20	0.0 ± 0.2	6 ± 34	0 ± 13	0.0 ± 0.3	1 ± 7
		Single-scale perceptual loss (layer 7)	-22 ± 40*	-17 ± 38*	-0.3 ± 0.8*	-17 ± 36*	-0.4 ± 0.9*	-36 ± 74*	-11 ± 21*	-0.1 ± 0.2*	-12 ± 34*	-4 ± 14*	-0.2 ± 0.4*	-4 ± 8*
		Multi-scale perceptual loss	-14 ± 39*	-10 ± 36	-0.3 ± 0.8*	-11 ± 35	-0.3 ± 0.9*	-29 ± 74*	-8 ± 20*	-0.0 ± 0.2	-7 ± 33	-2 ± 13	-0.1 ± 0.3	-2 ± 7
		Weighted multi-scale perceptual loss	-13 ± 40	-9 ± 39	-0.2 ± 0.8	-9 ± 37	-0.3 ± 0.9*	-31 ± 75*	-8 ± 21*	0.0 ± 0.2	-5 ± 35	-2 ± 14	-0.1 ± 0.3	-2 ± 8

Table 2. Reference dose values, dose uncertainties (MAE) and systematic dose uncertainties (ME) for each pseudo-CT generation method for each volume of interest

The mean values of DHV points are reported for the reference CT. The dose uncertainty is defined as the mean absolute DVH differences between the DVH calculated from the reference CT and those obtained from the pCTs. The systematic dose uncertainty is defined as the mean DVH differences between the DVH calculated from the reference CT and those obtained from the pCTs. The Wilcoxon test was used to: firstly, compare the dose uncertainty (MAE) of the GAN with L2 loss to those of the other methods; and secondly, to compare the systematic dose uncertainty (ME) of the methods to a null distribution. Significant differences ($p \leq 0.05$) are displayed using the symbol *. The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.

			Gamma pass-rate (%)	Mean gamma	Gamma pass-rate (%)	Mean gamma
			1%/1 mm, 10% low dose threshold		1%/1 mm, 30% low dose threshold	
Methods used to generate pseudo-CT	Patch-based method		98.7 ± 1.4*	0.47 ± 0.20*	99.5 ± 1.3	0.40 ± 0.16*
	U-Net methods	L2 loss	99.2 ± 1.0	0.39 ± 0.17	99.5 ± 1.5	0.33 ± 0.19
		Single-scale perceptual loss (layer 7)	99.3 ± 0.8	0.42 ± 0.13* °	99.8 ± 0.6	0.37 ± 0.15*
	GAN methods	L2 loss	99.1 ± 1.0	0.39 ± 0.16	99.6 ± 1.3	0.32 ± 0.18
		Single-scale perceptual loss (layer 7)	99.3 ± 0.9*	0.41 ± 0.15	99.7 ± 0.9	0.38 ± 0.16*
		Multi-scale perceptual loss	99.2 ± 0.8	0.40 ± 0.14	99.7 ± 0.9	0.35 ± 0.15
		Weighted multi-scale perceptual loss	99.3 ± 0.8*	0.40 ± 0.13	99.6 ± 1.1	0.36 ± 0.16*

Table 3. Mean gamma and gamma pass-rate calculated from the reference CT and pseudo-CT dose distributions according to each method

Values are mean ± standard deviation.

The Wilcoxon test was used to compare the gamma values of the GAN with L2 loss to those of the other methods. Significant differences ($p \leq 0.05$) are displayed using the symbol *.

The Wilcoxon test was also used to compare the gamma values of the GAN with perceptual loss to those of the U-Net with perceptual loss. Significant differences ($p \leq 0.05$) are displayed using the symbol °.