International Journal of Radiation Oncology biology • physics

www.redjournal.org

Physics Contribution

Dosimetry-Driven Quality Measure of Brain Pseudo Computed Tomography Generated From Deep Learning for MRI-Only Radiation Therapy Treatment Planning

Emilie Alvarez Andres, MSc,[*,†,‡] Lucas Fidon, MSc,[†,§] Maria Vakalopoulou, PhD,[§] Marvin Lerousseau, MSc,[*,‡,§] Alexandre Carré, MSc,[*,‡] Roger Sun, MD,[*,‡,§,‖] Guillaume Klausner, MD,[*,‖] Samy Ammari, MD,[¶] Nathan Benzazon, MSc,[*,‡] Sylvain Reuzé, PhD,[*,‡] Théo Estienne, MSc,[*,‡,§] Stéphane Niyoteka, MSc,[*,‡] Enzo Battistella, MSc,[*,‡,§] Angéla Rouyar, PhD,[*,‡] Georges Noël, MD, PhD,[#] Anne Beaudre, PhD,[‡] Frédéric Dhermain, MD, PhD,[‖] Eric Deutsch, MD, PhD,[*,‖] Nikos Paragios, PhD,[†] and Charlotte Robert, PhD[*,‡]

U1030 Molecular Radiotherapy, Paris-Sud University - Gustave Roussy - Inserm - Paris-Saclay University, Villejuif, France; [†]TheraPanacea, Paris, France; [‡]Department of Medical Physics, Gustave Roussy - Paris-Saclay University, Villejuif, France; [§]MICS Laboratory, CentraleSupélec, Paris-Saclay University, Gif-sur-Yvette, France; [‖]Department of Radiotherapy, Gustave Roussy - Paris-Saclay University, Villejuif, France; [¶]Department of Radiology, Gustave Roussy - Paris-Saclay University, Villejuif, France; and [#]Department of Radiotherapy, Paul Strauss Institute, Strasbourg, France

Received Nov 15, 2019. Accepted for publication May 5, 2020.

Purpose: This study aims to evaluate the impact of key parameters on the pseudo computed tomography (pCT) quality generated from magnetic resonance imaging (MRI) with a 3-dimensional (3D) convolutional neural network.

Methods and Materials: Four hundred two brain tumor cases were retrieved, yielding associations between 182 computed tomography (CT) and T1-weighted MRI (T1) scans, 180 CT and contrast-enhanced T1-weighted MRI (T1-Gd) scans, and

Corresponding author: Charlotte Robert, PhD; E-mail: ch.robert@gustaveroussy.fr

Emilie Alvarez Andres and Lucas Fidon made equal contributions to this study.

Disclosures: E.A.A. reports grants from TheraPanacea, during the conduct of the study. S.R. has been a full-time employee of GE Healthcare since December 2018, outside of the submitted work.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 880314. E.D. reports grants and personal fees from Roche Genentech, grants from Servier, grants from Astrazeneca, grants and personal fees from Merck Serono, grants from BMS, and grants from MSD, outside the submitted work. N.P. was CEO and founder of TheraPanacea, during the conduct of the study.

Research data: Research data are not available at this time.

Supplementary material for this article can be found at https://doi.org/10.1016/j.ijrobp.2020.05.006.

Int J Radiation Oncol Biol Phys, Vol. 108, No. 3, pp. 813–823, 2020
0360-3016/$ - see front matter
https://doi.org/10.1016/j.ijrobp.2020.05.006

40 CT, T1, and T1-Gd scans. A 3D CNN was used to map T1 or T1-Gd onto CT scans and evaluate the importance of different components. First, the training set size's influence on testing set accuracy was assessed. Moreover, we evaluated the MRI sequence impact, using T1-only and T1-Gd−only cohorts. We then investigated 4 MRI standardization approaches (histogram-based, zero-mean/unit-variance, white stripe, and no standardization) based on training, validation, and testing cohorts composed of 242, 81, and 79 patients cases, respectively, as well as a bias field correction influence. Finally, 2 networks, namely HighResNet and 3D UNet, were compared to evaluate the architecture's impact on the pCT quality. The mean absolute error, gamma indices, and dose-volume histograms were used as evaluation metrics.

Results: Generating models using all the available cases for training led to higher pCT quality. The T1 and T1-Gd models had a maximum difference in gamma index means of 0.07 percentage point. The mean absolute error obtained with white stripe was 78 ± 22 Hounsfield units, which slightly outperformed histogram-based, zero-mean/unit-variance, and no standardization ($P < .0001$). Regarding the network architectures, 3%/3 mm gamma indices of 99.83% ± 0.19% and 99.74% ± 0.24% were obtained for HighResNet and 3D UNet, respectively.

Conclusions: Our best pCTs were generated using more than 200 samples in the training data set. Training with T1 only and T1-Gd only did not significantly affect performance. Regardless of the preprocessing applied, the dosimetry quality remained equivalent and relevant for potential use in clinical practice.

Introduction

Magnetic resonance imaging (MRI) has become prevalent in radiation therapy planning owing to its excellent soft tissue contrast compared with computed tomography (CT). During the brain tumor radiation therapy process, MRI and CT play a key role in indicating areas of interest and estimating the dosimetry, respectively. Yet, dealing with multiple imaging modalities requires coregistration, leading to errors of up to 2 mm[1] and target volume margin increases.

To address this limitation, numerous approaches have been developed to generate a pseudo computed tomography (pCT) from MRI.[2,3] First, the bulk density approach[4,5] assigns specific electron densities (EDs) to presegmented MRI scans; however, this relies on the labeling quality. Second, the multiatlas method constitutes a multiple-atlas database representing coregistered pairs of CT and MRI acquired from different patients. The incoming MRI is first aligned to the atlas MRI through a deformable registration. The resulting deformation fields are then applied to the atlas CT scans, which are combined to generate the pCT.[6,7] Because of the computational complexity of deformable registrations, the multiatlas approach is time-consuming. To mitigate these limitations, deep learning (DL) methods[8-10] have been recently introduced, with promising results.[11,12] Compared with the other approaches, DL-based methods efficiently exploit large databases to learn direct mapping from MRI to CT.

A deep convolutional neural network (CNN) consists of a composition of convolutional filters and simple nonlinear functions organized in layers. The parameters of the CNN are learned using pairs of MRI/CT training data via empirical risk minimization and stochastic gradient descent. DL-based methods benefit from highly efficient graphical processing unit implementations, which reduce the inference time of the pCT by several orders of magnitude compared with atlas-based methods. Based on an NVIDIA Titan X graphical processing unit, Han et al[13] reported durations of 9 seconds and 10 minutes for the DL and atlas-based approaches, respectively.

However, there is still no consensus regarding (1) the optimal training set size, (2) the best-suited magnetic resonance (MR) sequence, (3) the optimal MR standardization preprocessing, (4) the use of an inhomogeneity correction, and (5) the best suited network architecture (Table EA1). Additionally, there has been no discussion about the approach to evaluate the generated pCT.

Indeed, training data sets sizes ranging from 15[14] to 77 patients[12] have been reported, raising the issue of the minimal number of training patients required to ensure a satisfying generalization to unseen examples. Moreover, most of the studies used either T1-weighted MRI (T1) or contrast-enhanced T1-weighted MRI (T1-Gd). However, the benefit of using a contrast agent in terms of pCT quality is still unclear. Additionally, only a few studies have applied MRI intensity standardization as preprocessing. Yet, doing so can improve pCT quality.[15] A similar question concerns bias field correction; only Han et al[13] have applied it. Finally, several CNN architectures have been used in the literature, such as HighResNet[16,17] and UNet,[13] with no systematic comparisons. An additional aspect not explicitly discussed in these works is the influence of these parameters on dosimetry-based pCT evaluation. Numerous studies report their performance using peak signal-to-noise ratio or mean absolute error (MAE) metrics,[13,18,19] which may be irrelevant to the real clinical scenario.

This study aims to evaluate the impact of significant parameters, namely the training data set size, input MR sequence, standardization strategy, application of inhomogeneity correction, and network architecture, on the computed pCT's accuracy and the associated clinical dosimetry. The pCT evaluation is based on both the MAE (based on the intensities and ED) and clinical criteria, namely 1%/1 mm, 2%/2 mm, and 3%/3 mm gamma indices and differences in dose-volume histograms (DVHs) of the planning target volume.

Methods and Materials

Images acquisition and preprocessing

Four hundred two institutional patients treated between 2006 and 2017 for brain tumors were included in this retrospective study. For all patients, the delay between the planning CT and T1 or T1-Gd MR acquisitions did not exceed 8 days. The data set was composed of 182 CT/T1, 180 CT/T1-Gd, and 40 CT/T1/T1-Gd paired images.

All CT scans were acquired with a Sensation Open scanner (Siemens Healthineers, Erlangen, Germany) using a 120kVp tube voltage. The slice thickness was equal to 1 mm, 2 mm, 3 mm, and 5 mm for 3, 45, 353, and 1 patient cases, respectively. The native X and Y voxel sizes were included in (0.50 mm; 0.70 mm), (0.70 mm; 0.90 mm), and (0.90 mm; 1.10 mm) for 208, 76, and 118 patients, respectively.

The MRI scans were all acquired with GE Healthcare devices (GE Healthcare, Milwaukee, WI). Two patients' MR sequences were from external institutes and were acquired on 2 different 1.5T devices: Optima MR360 and Discovery MR450. The remaining MRI scans were institutional images, acquired on a 3T Discovery MR750w (224 patient cases), a 1.5T Optima MR450w (9 patient cases), or a 1.5T Signa Excite (167 patient cases). Only 3-dimensional (3D) axial T1-weighted images with or without a gadolinium injection were used. Initial slice thicknesses were included in (1 mm; 1.2 mm), (1.4 mm; 2 mm), (3 mm; 3.2 mm), and equal to 5 mm for 234, 10, 157, 1 patient, respectively. Regarding the native X and Y voxel sizes, they were included in (0.44 mm; 0.50 mm), (0.50 mm; 0.58 mm), and equal to 0.94 mm for 325, 73 and 4 patients, respectively.

For each patient, the CT was first rigidly registered to the T1 or T1-Gd images using the Drop library (https://github.com/biomedia-mira/drop2). The images then were linearly resampled to a 1 mm $\times$ 1 mm $\times$ 1 mm voxel size, before harmonizing the volumes to 300 $\times$ 300 $\times$ 242 voxels. Both the MRI intensities and the CT Hounsfield units were clipped, to 0.1 and 99.9 percentiles and ($-$1000 HU, 1800 HU), respectively. The maximum HU was empirically determined based on CT intensity histograms. Finally, the Hounsfield units were rescaled between ($-$1, 1).

Lastly, 60%, 20%, and 20% of the patients were randomly parsed into training, validation, and testing sets, provided that the T1 and T1-Gd were equal in proportion. Patients with all CT, T1, and T1-Gd images were automatically assigned to the testing set to be used for the dosimetry-based evaluation.

Standardization strategies

Three approaches were adopted to standardize the MRI.

The first approach was a histogram-based standardization (HB) based on the method described by Nyúl et al.[20] HB consists of matching percentiles (10, 20, 30, 40, 50, 60, 80, 90) of an image to predefined template values that are computed using the MR images of the cohort. The intensity match is obtained via a piecewise linear transformation applied to image intensities.

The second approach consists of a normalization of the intensity distribution inside the head of each patient to zero mean and unit variance (ZMUV).[15]

The last method, white stripe (WS),[21] is similar to the ZMUV approach but is based on the normal-appearing white matter mean and standard deviation, because it is known to be homogeneous. Brain masks were first extracted with the HD-BET tool.[22] The MR images were then normalized with the intensity-normalization package.[15]

Network architectures

Following popular choices of network architectures in the literature, we decided to use the HighResNet 3D CNN presented by Li et al[23] and the 3D UNet.[24]

The HighResNet was originally designed for a segmentation task. In contrast to other networks, it preserves the image resolution (no pooling layers) and is compact (0.8 million parameters). The main components of the network were the dilated 3D convolutions with kernels of size 3 $\times$ 3 $\times$ 3, the residual connections, the normalization layers, and the Rectified Linear Unit (ReLU) activations. These operations were organized into 9 residual blocks based on convolution filter sizes dilated by 1, 2 or 4. Each block contained a series of normalization, ReLU, and convolution, which was repeated twice before adding the block input to its output. The 2 final layers were not residual and were composed of 3 $\times$ 3 $\times$ 3 and 1 $\times$ 1 $\times$ 1 convolutions to obtain the final pCT volume.

The 3D UNet is a popular encoder$-$decoder neural network architecture in medical image computing. It is characterized by its long shortcut connections between layers output at different stages of the network architecture, which give it a U-shape. These connections allow it to combine features at different scales and different spatial resolutions. Contrary to the HighResNet, 3D UNet uses max-pooling layers and no dilated convolutions. This difference enables the 3D UNet to have more features and to use larger input patches than the HighResNet at the price of a lower spatial resolution in some layers of the 3D UNet. ReLU activation, 3 $\times$ 3 $\times$ 3 convolutions, instance normalization, and linear upsampling were used for the 3D UNet, resulting in approximately 15 million parameters.

The final aim of this work was not to develop an original network but to provide guidelines for future pCT studies by evaluating the impact of different parameters on the pCT quality in terms of image intensity and dosimetry. As a result, we adapted the HighResNet for pCT generation. We replaced the normalization layers with instance normalization,[25] removed the softmax layer after the last convolutional layer, and changed the output channel number to

1. The modified network architecture is displayed in Figure EA2.

To optimize the network parameters, we used the MAE loss function:

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} \left| I_{CT}(i) - I_{pCT}(i) \right| \quad (1)$$

where $I_{CT}(i)$ and $I_{pCT}(i)$ are the intensities of the CT and the pCT at voxel i, and N is the considered number of voxels.

Owing to memory constraints, patches of size 96 × 96 × 96 voxels and 136 × 136 × 136 voxels were used as input of the HighResNet and 3D UNet, respectively. At inference, the 3D MRI scans were divided into patches to reconstruct the whole pCT. A patch margin of length 5 and 1 voxels for the HighResNet and 3D UNet, respectively, was applied, leading to predictions inside subpatches of size 86 × 86 × 86 and 134 × 134 × 134. The motivation of the margins is to guarantee a smooth transition between patches prediction. Note that patches overlapped, contrary to subpatches. The overlap process is described in Figure EA3.

For both networks, the learning rate was set to 0.001. Early stopping on the validation set was used as stopping criterion to assess the convergence of the CNN. Dropout was used after the penultimate layer during training with a probability of 0.5.

Note that no data augmentation was used in this study.

Impact of key parameters

The first experiment consisted of quantifying the impact of the training set size. Five different HighResNet networks were trained using 242 (121 T1, 121 T1-Gd), 121 (61 T1, 60 T1-Gd), 60 (30 T1, 30 T1-Gd), 30 (15 T1, 15 T1-Gd), and 15 (8 T1, 7 T1-Gd) patients, respectively, in the training set. The validation and testing cohorts were the same for all the training set sizes and included 81 (41 T1, 40 T1-Gd) and 79 (39 T1, 40 T1-Gd) cases, respectively. All the MR images were standardized using the HB method.

A second experiment was conducted to determine the T1 input sequence best suited to generate the pCT. We constituted 2 HB-standardized cohorts: (1) a T1-only cohort with 134, 44, and 40 T1 MRI cases for the training, validation, and testing sets, respectively, and (2) a T1-Gd−only cohort with 133, 44, and 40 cases, respectively. The cases included in the 2 testing cohorts were the same, for a fair comparison. For this experiment, different T1 and T1-Gd histogram templates were computed for the HB standardization, based on the 134 and 133 patients included in the training cohorts. Experiment 2 was based on the HighResNet.

The third experiment assessed the role of the MRI standardization using 242 (121 T1, 121 T1-Gd), 81 (41 T1, 40 T1-Gd), and 79 (39 T1, 40 T1-Gd) cases in the training, validation, and testing sets, respectively. The HighResNet architecture was used for this experiment. Four different approaches were investigated: HB, ZMUV, WS, and no standardization (NS).

The fourth experiment was performed to evaluate the role of the bias field correction, using HighResNet. As a result, the N4 filter[26] was optionally applied on MR images. The best standardization technique defined by experiment 3 was used here. The training, validation, and testing sets were those used in experiment 3.

The last experiment was conducted to analyze the influence of the network architecture on the quality of the generated pCT. To this aim, the HighResNet used in the previous experiments and the 3D UNet were trained, validated, and tested. The best preprocessing strategies highlighted by the third and fourth experiments were applied. The split of the data set was the same as experiment 3.

A summary of the experiments is presented in Figure EA4.

Evaluation criteria

First, the initial CT and the pCT were compared using the MAE (Equation 1). It was computed in 4 different areas: whole head, air, bone, and water. The head was segmented using the Otsu approach.[27] The other regions were obtained thresholding the CT: $x \leq -200$ *HU*, -200 *HU* $< x <$250 *HU*, and 250 HU $\leq x$ for the air, water, and bone regions, respectively. The MAE was calculated from the 3D intensity volumes or the 3D ED volumes obtained applying the HU-ED calibration curve.

Furthermore, we evaluated the pCT quality in terms of dose prediction for all the experiments, except the first one, by computing metrics used in clinics: 1%/1 mm, 2%/2 mm, and 3%/3 mm 3D global gamma indices were considered, and no dose threshold was applied. In addition, relative differences between CT and pCT DVH ($D_{02\%}$, $D_{50\%}$, $D_{95\%}$, and $D_{98\%}$) of the planning target volume were calculated.

The dosimetry plans from the original CT were recalculated on the pCT, with the pencil beam (PB) dose calculation algorithm implemented in iPlan RT 4.5 Dose (Brainlab, Munich, Germany).[28] The default grid size was set to 2 mm. It is worth noting the grid was adaptive, meaning that it became finer for small objects. This approach was combined with a ray-tracing technique that was applied during the radiologic path length calculation. These 2 approaches sped up the dose calculation. For this dosimetry analysis, a subset cohort of the testing set, corresponding to cases whose dosimetry had been realized with iPlan, was used. It was composed of 39 grades 3 and 4 glioma cases (19 T1, 20 T1-Gd) treated with a sliding window intensity modulated radiation therapy approach, delivered with a 6 MV beam. A total of 18, 11, 7, 2, and 1 patient cases were treated with 5, 6, 7, 8, and 10 beams, respectively. An illustration of the overall workflow is presented in Figure EA5.

Two-sided paired Wilcoxon tests, with a significance level set to .05, were performed as statistical analysis. Only results computed on the testing set are reported.

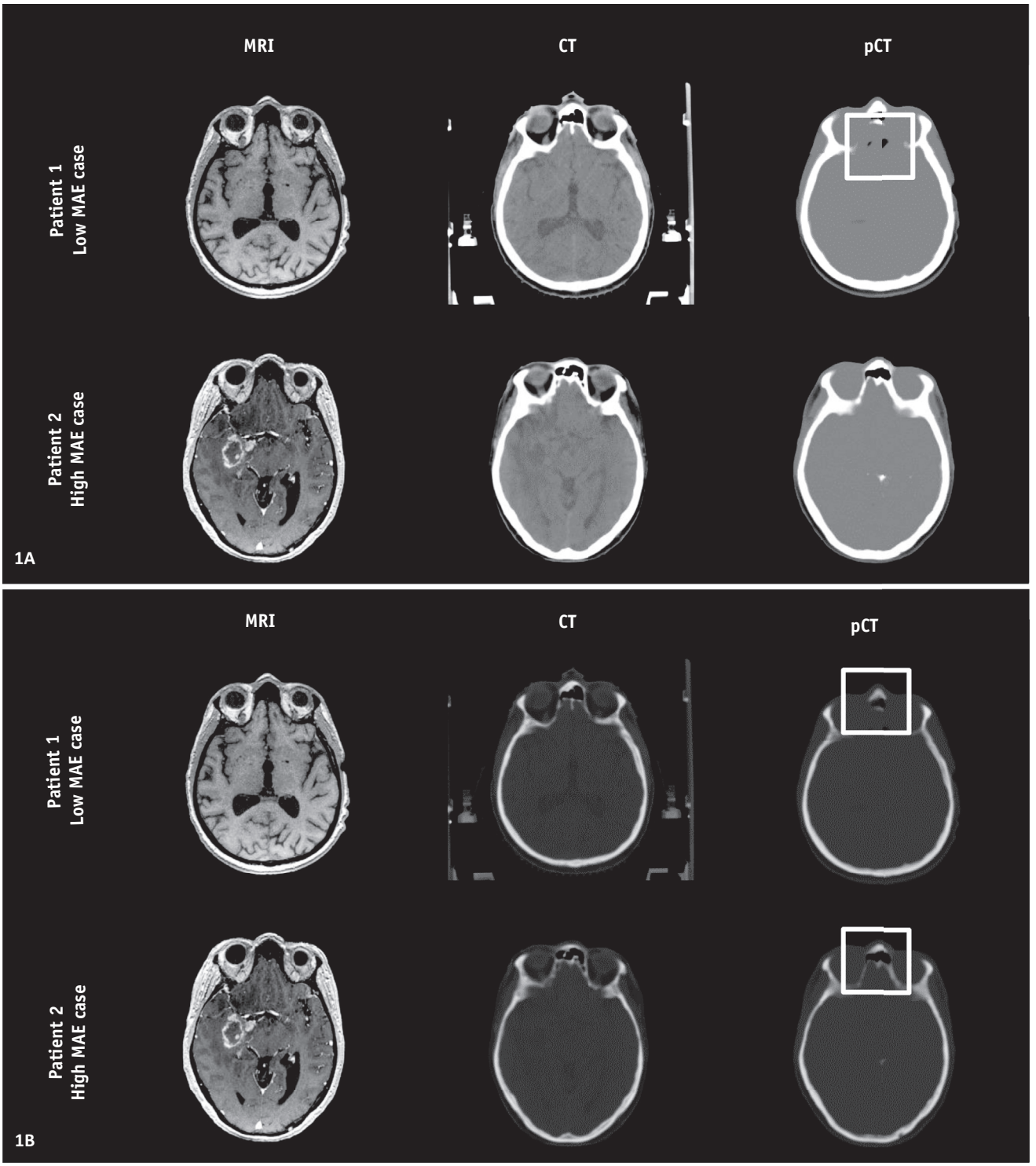

Fig. 1. (From left to right) magnetic resonance imaging, original computed tomography, and pseudo computed tomography with soft tissue (A) and bone (B) windows widths and levels, respectively, for 2 patients. The squares highlight some of the incorrect reconstructed areas.

Results

Figures 1A and 1B present examples of MRI, CT, and pCT with soft tissue and bone windows widths and levels, respectively. They were extracted from the third experiment, using the HighResNet and the HB intensities standardization. The first line corresponds to a low MAE case (head MAE = 64 HU) and the second line to a high MAE

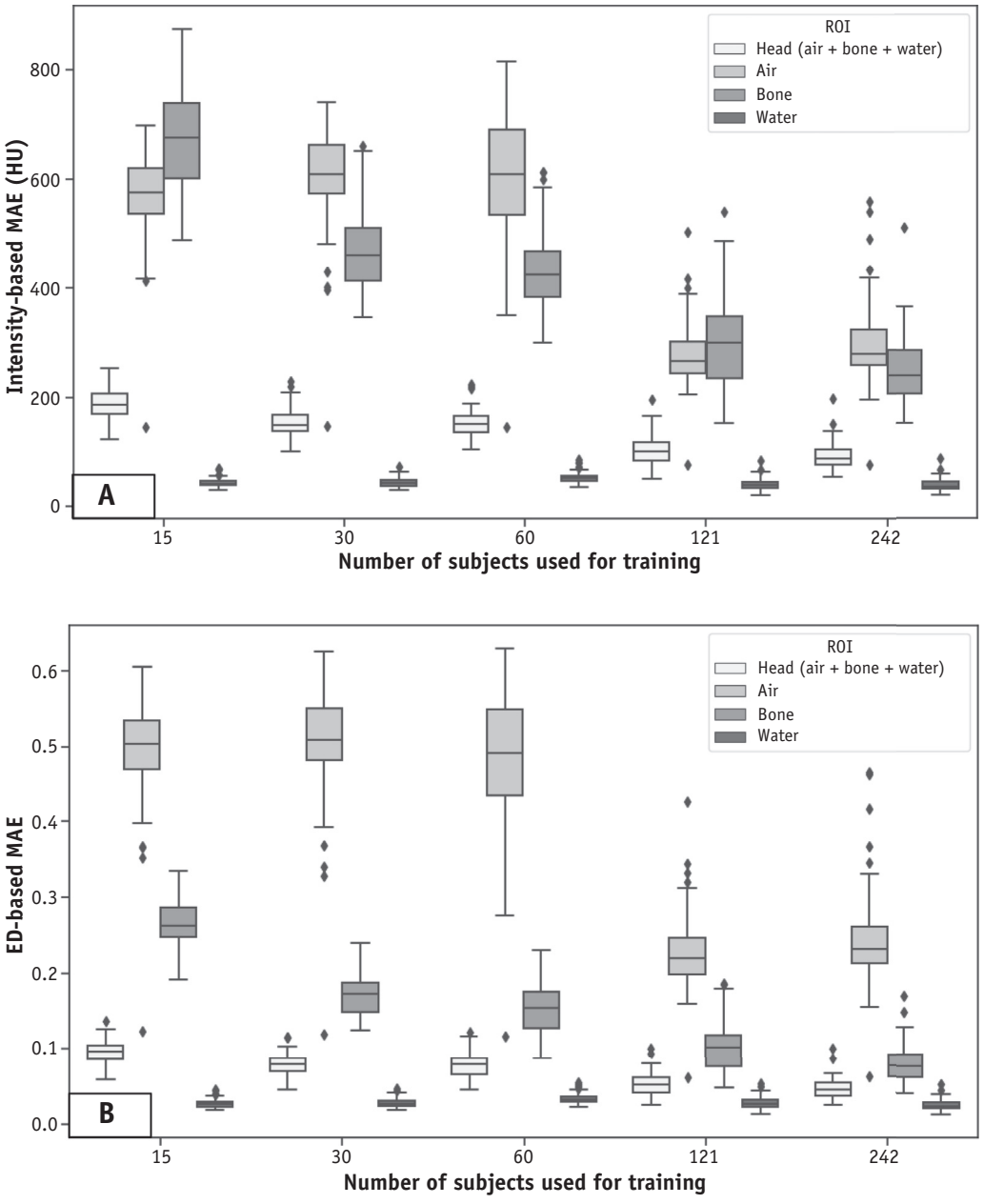

Fig. 2. Evolution of the mean absolute error (MAE) based on Hounsfield units (A) and electron densities (B) when modifying the number of training subjects. The boxplot corresponds to the first and third MAE quartiles with the MAE median in the middle; the whiskers correspond to the range of the MAE distribution after excluding the outliers.

case (head MAE = 110 HU). Some air and bone areas appear to be less accurately reconstructed, as highlighted by the squares.

The intensity-based MAE obtained from different training set sizes is displayed in Figure 2A. For the head area, increasing the training data set resulted in a decrease of the MAE mean ± standard deviation (std) from 189 ± 28 HU for the 15-patient training set model to 92 ± 23 HU corresponding to the 242-patient training set model. Bone and air regions had the highest MAE. Differences between all the training size models were significant for the head region ($P < .0001$) except between 30 and 60 patients (Table EA6).

The ED-based MAE is presented in Figure 2B, to more accurately assess the pCT quality with respect to its clinical use. A similar behavior is observed, with a head MAE decrease from 0.10 ± 0.01 to 0.05 ± 0.01 when increasing the training set size from 15 to 242.

Table 1 presents the mean ± std of the MAE, gamma index, DVH difference, and Wilcoxon test values derived from the T1-only and T1-Gd–only models. The maximum differences between the T1 and T1-Gd models obtained for the head MAE means and gamma index means were equal to 3 HU and 0.07 percentage points (pp), respectively.

Mean ±std of the MAE, gamma indices, and DVH differences obtained for the standardization experiment are provided

Table 1 Means ± standard deviations of the MAE, gamma indices, and DVH differences computed for the PTV and statistical analysis derived from the T1-weighted MRI (T1) and contrast-enhanced T1-weighted MRI (T1-Gd) cohort comparison

	T1 only	T1-Gd only	*P* value	95% confidence interval
MAE head, HU	84 ± 25	87 ± 28	**.0047**	−3.93 to −0.76
MAE air, HU	274 ± 63	306 ± 74	**<.0001**	−36.51 to −22.37
MAE bone, HU	228 ± 63	236 ± 71	.066	−11.38 to 0.48
MAE water, HU	38 ± 11	38 ± 12	.82	−0.83 to 0.73
1%/1 mm gamma index	97.87% ± 1.16%	97.94% ± 1.07%	.59	−0.12 to 0.05
2%/2 mm gamma index	99.60% ± 0.33%	99.63% ± 0.30%	.50	−0.05 to 0.02
3%/3 mm gamma index	99.84% ± 0.18%	99.85% ± 0.18%	.44	−0.03 to 0.01
Difference PTV $D_{02\%}$	0.20% ± 0.15%	0.15% ± 0.09%	**.0041**	0.02-0.08
Difference PTV $D_{50\%}$	0.20% ± 0.15%	0.13% ± 0.08%	**.015**	0.02-0.12
Difference PTV $D_{95\%}$	0.20% ± 0.17%	0.14% ± 0.10	**.012**	0.02-0.12
Difference PTV $D_{98\%}$	0.27% ± 0.37%	0.22% ± 0.41%	**.026**	0.01-0.12

Abbreviations: DVH = dose-volume histogram; MAE = mean absolute error; PTV = planning target volume.
Bolded *P* values are associated with significant distributions differences.

in Table 2. The statistical analysis is presented in Table EA7. WS led to a head MAE of 78 ± 22 HU, which was significantly lower than the 3 other methods ($P < .0001$). Regarding the dosimetry, 3%/3 mm gamma indices of 99.86% ± 0.16%, 99.83% ± 0.19%, 99.85% ± 0.17%, and 99.86% ±0.18% were achieved for the HB, ZMUV, WS, and NS approaches.

Regarding the fourth experiment based on the combination of the HighResNet with the WS standardization, mean ± std of the MAE and dosimetry metrics are presented in Table 3. Applying the bias field correction led to a head MAE of 81 ± 22 HU. Concerning the DVH D02%, differences equal to 0.15% ± 0.12% and 0.20% ± 0.13% were achieved with and without the application of the N4 filter, respectively ($P = .026$).

Table 4 provides the MAE and dosimetry values for the last experiment, which was conducted to compare the HighResNet with the 3D UNet. For both networks, the WS MRI standardization and the N4 filter were applied. The mean ± std obtained for the head MAE was 81 ± 22 HU and 90 ± 21 HU for the HighResNet and 3D UNet, respectively ($P < .0001$). Significantly higher gamma indices were obtained with the HighResNet ($P < .0001$), with a pass rate of 97.92% ± 1.06% for the most restrictive 1%/1 mm criterion.

Discussion

This study aimed at evaluating the impact of key parameters of brain pCT generation from T1 or T1-Gd images,

Table 2 Means ± standard deviations of the MAE, gamma indices, and DVH differences computed for the PTV derived from the HB, ZMUV, WS, and NS cohorts

	HB	ZMUV	WS	NS
MAE head, HU	92 ± 23	83 ± 22	78 ± 22	96 ± 23
MAE air, HU	297 ± 73	284 ± 62	253 ± 65	313 ± 68
MAE bone, HU	251 ± 61	214 ± 55	199 ± 54	252 ± 60
MAE water, HU	39 ± 11	38 ± 12	36 ± 11	43 ± 11
1%/1 mm gamma index	97.94% ± 1.06%	97.90% ± 1.10%	98.08% ± 1.01%	97.80% ± 1.17%
2%/2 mm gamma index	99.63% ± 0.28%	99.61% ± 0.30%	99.64% ± 0.29%	99.61% ± 0.31%
3%/3 mm gamma index	99.86% ± 0.16%	99.83% ± 0.19%	99.85% ± 0.17%	99.86% ± 0.18%
Difference PTV $D_{02\%}$	0.22% ± 0.17%	0.22% ± 0.16%	0.20% ± 0.13%	0.24% ± 0.20%
Difference PTV $D_{50\%}$	0.24% ± 0.16%	0.23% ± 0.16%	0.21% ± 0.13%	0.27% ± 0.17%
Difference PTV $D_{95\%}$	0.27% ± 0.31%	0.21% ± 0.17%	0.19% ± 0.15%	0.32% ± 0.32%
Difference PTV DVH $D_{98\%}$	0.38% ± 0.58%	0.27% ± 0.35%	0.20% ± 0.17%	0.38% ± 0.46%

Abbreviations: DVH = dose-volume histogram; HB = histogram-based; MAE = mean absolute error; NS = no standardization; PTV = planning target volume; WS = white stripe; ZMUV = zero mean/unit variance.

Table 3 Mean ± standard deviations of the MAE, gamma indices, and DVH differences of the PTV and statistical analysis derived from the WS and WS combined with a bias field correction (N4) cohort comparison

	WS	WS and N4	*P* value	95% confidence interval
MAE head, HU	78 ± 22	81 ± 22	**<.0001**	−4.79 to −2.57
MAE air, HU	253 ± 65	244 ± 62	**<.0001**	5.23-11.84
MAE bone, HU	199 ± 54	230 ± 56	**<.0001**	−35.81 to −27.07
MAE water, HU	36 ± 11	34 ± 10	**<.0001**	2.02-2.91
1%/1 mm gamma index	98.08% ± 1.01%	97.92% ± 1.06%	**.0035**	0.04-0.19
2%/2 mm gamma index	99.64% ± 0.29%	99.60% ± 0.32%	**.0026**	0.01-0.06
3%/3 mm gamma index	99.85% ± 0.17%	99.83% ± 0.19%	**.012**	0.00-0.03
Difference PTV $D_{02\%}$	0.20% ± 0.13%	0.15% ± 0.12%	**.026**	0.00-0.13
Difference PTV $D_{50\%}$	0.21% ± 0.13%	0.13% ± 0.10%	**.0017**	0.03-0.15
Difference PTV $D_{95\%}$	0.19% ± 0.15%	0.11% ± 0.12%	**.0034**	0.03-0.14
Difference PTV $D_{98\%}$	0.20% ± 0.17%	0.13% ± 0.13%	**.0088**	0.02-0.14

Abbreviations: DVH = dose-volume histogram; MAE = mean absolute error; PTV = planning target volume; WS = white stripe.
Bolded *P* values are associated with significant distributions differences.

namely the training set size, the MR input sequence, the standardization strategy, the application of a bias field correction, and the network architecture. Best results were achieved when combining the WS MRI standardization with an inhomogeneity correction, the HighResNet, and all our 242 training patient cases. This suggests that more training cases could lead to further improvements.

Regarding the MR sequences experiment, a difference of 3 HU was observed between the head MAE means of the T1 only and T1-Gd−only models, suggesting that the contrast agent resulted in a negligible pCT improvement. The DVH differences led to a similar conclusion, as only a 0.07 pp maximum difference between the 2 models means was obtained. We conducted an extra experiment to evaluate the potential benefit of the T2 fluid attenuated inversion recovery (FLAIR) MR sequence. A total of 134, 44, and 40 patients were included in the training, validation, and testing sets, respectively. The preprocessing described for the T1-only and T1-Gd−only cohorts was similarly applied. A mean MAE ± std of 115 ± 22 HU was obtained within the head area. Differences with the T1-only and T1-Gd−only cohorts were found to be significant ($P <$ .001). Thus, T2 FLAIR appeared to generate largest pCT intensity-linked errors. It could be attributed to the lower contrast contained in T2 FLAIR images compared with T1/T1-Gd images. A second interpretation could be the slice thickness, which was larger for most of the T2 FLAIR images compared with T1/T1-Gd images, resulting in a less

Table 4 Mean ± standard deviations of the MAE, gamma indices, and DVH differences computed for the PTV and statistical analysis derived from WS combined with a bias field correction (N4) and initial HighResNet against WS associated with N4 and 3D UNet cohort comparison

	WS, N4, and HighResNet	WS, N4, and 3D UNet	*P* value	95% confidence interval
MAE head, HU	81 ± 22	90 ± 21	**<.0001**	−9.39, −6.99
MAE air, HU	244 ± 62	266 ± 66	**<.0001**	−27.18 to −15.56
MAE bone, HU	230 ± 56	209 ± 54	**<.0001**	16.91-25.79
MAE water, HU	34 ± 10	49 ± 11	**<.0001**	−15.81 to −14.09
1%/1 mm gamma index	97.92% ± 1.06%	97.28% ± 1.46%	**<.0001**	0.42-0.79
2%/2 mm gamma index	99.60% ± 0.32%	99.39% ± 0.47%	**<.0001**	0.10-0.24
3%/3 mm gamma index	99.83% ± 0.19%	99.74% ± 0.24%	**<.0001**	0.03-0.11
Difference PTV $D_{02\%}$	0.15% ± 0.12%	0.33% ± 0.21%	**<.0001**	−0.28 to −0.11
Difference PTV $D_{50\%}$	0.13% ± 0.10%	0.29% ± 0.19%	**<.0001**	−0.22 to −0.10
Difference PTV $D_{95\%}$	0.11% ± 0.12%	0.28% ± 0.18%	**<.0001**	−0.24 to −0.13
Difference PTV $D_{98\%}$	0.13% ± 0.13%	0.31% ± 0.18%	**<.0001**	−0.26 to −0.15

Bolded *P* values are associated with significant distributions differences.

informative spatial sampling. Future work includes the comparison of T1 and unusual sequences, such as zero echo time in which bone areas are more visible, to assess which combination of MRI sequences is optimal for an accurate pCT reconstruction in radiation therapy.

The third experiment concerned the MRI standardization and used the HighResNet as network architecture. A mean ± std of 78 ± 22 HU was obtained for the head MAE when applying the WS standardization, which slightly outperformed HB, ZMUV, and NS ($P < .0001$). The greatest errors were located in the air and bone areas, with a MAE of 253 ± 65 HU and 199 ± 54 HU, respectively, and seemed to correspond to misaligned regions or areas with high dose gradients.

Dinkla et al[11] reported competitive head MAE of 67 ± 11 HU. All the CT and MR images used in their study were acquired on the same device. In this work, MR images were acquired from 5 different devices. Table EA8 presents the composition of the training, validation, and test sets in terms of MR devices. As one can notice, most of the MRI scans in the training set (133) were acquired with the DISCOVERY MR750w - 3T device. To analyze the impact of this unbalance, the test set was split into 2 subsets: MRI from the DISCOVERY MR750w 3T (57 patients) and MRI from the SIGNA EXCITE 1.5T (21 patients). The default HB standardization and HighResNet were used for this experiment. Mean head MAE ± std was 86 ± 22 HU for the DISCOVERY MR750w 3T and 106 ± 16 HU for the SIGNA EXCITE 1.5T ($P < .0001$). The pCT computed from the DISCOVERY MR750w 3T device were of higher quality because more MRI scans acquired with this device were included in the training set and because 3T devices offer better image resolution. Thus, we think that the composition of the training set had a nonnegligible impact on the generated pCT. Comparing the literature MAE is, however, not a trivial task owing to the use of heterogeneous data sets, suggesting the need for publicly available data sets.

Concerning the dosimetry analysis, negligible differences were observed between the different standardization approaches. Regarding WS, a mean ± std of 99.85% ± 0.17% was obtained for the 3%/3 mm gamma index, which was not significantly different from the ZMUV, HB, and NS gamma indices ($P \geq .14$). These nonsignificant dosimetry results can be attributed to the nonlinearity of both the HU-ED curve and the radiation matter interaction effects. Very few studies reported dosimetry evaluations for brain pCT generated with a DL-based approach. Dinkla et al[11] achieved 91.1% ± 3.0%, 95.8% ± 2.1%, and 99.3% ± 0.4% for 1%/1 mm, 2%/2 mm, and 3%/3 mm head gamma indices with no threshold. A similar performance was obtained by Liu et al,[29] who reported 99.2% for the 3%/3 mm gamma index. Recently, Kazemifar et al[12] achieved state of-the-art 1%/1 mm and 2%/2 mm gamma indices of 94.6% ± 2.9% and 99.2% ± 0.8%. Dosimetry analyses are crucial because they are the only relevant metric for use in clinics.

The fourth experiment evaluated the role of an inhomogeneity correction combined with the HighResNet and the WS standardization. Although a slight increase of 3 HU in the mean head MAE was obtained when applying the N4 filter, the DVH analysis showed a negligible decrease in the mean of up to 0.08 pp ($P \leq .026$). This could be justified by acceptable MRI quality or the network's ability to handle this issue.

The last experiment was the evaluation of 2 different network architectures: HighResNet and 3D UNet. For each model, the WS standardization and the N4 filter were applied. The mean head MAE ± std was 81 ± 22 HU and 90 ± 21 HU for the HighResNet and 3D UNet, respectively. The lower HighResNet MAE may be attributed to 2 major advantages: the dilated convolution filters, which enable a large spatial context while retaining the full image resolution, and the residual connections, which regularize the optimization of the model.

Regarding the dosimetry, 3%/3 mm gamma indices equal to 99.83% ± 0.19% and 99.74% ± 0.24% were obtained for the HighResNet and the 3D UNet, respectively. As a result, no significant clinical impact was observed between the 2 architectures. In the literature, a lower MAE of 47 ± 11 HU was reported by Kazemifar et al[12] using a 2D GAN. In the context of pCT generation, a GAN corresponds to the training of a second auxiliary neural network that learns a loss function to estimate the distance between a pCT and the distribution of all the true possible CT. This data-driven loss function is used to train the main neural network that learns the mapping from MRI to pCT. Therefore, pCTs produced by a GAN are not guaranteed to respect the anatomy of the patient. To mitigate this issue, CycleGAN using an additional cycle-consistency penalization has been proposed.[19,30] However, cycle consistency implies a 1-to-1 mapping between the MRI and CT, which is not realistic and can lead to artefacts in the pCT.[31] As a result, further investigation of the errors specific to GAN and CycleGAN is needed for their clinical use in radiation therapy and is beyond the scope of this paper.

The loss function used to train the network has a knock-on effect on the pCT quality. Here, the MAE was chosen because it was found to generate less blurry images than the mean squared error during preliminary experiments. Kazemifar et al[12] trained two 2D GAN based on the MAE and the mutual information loss functions and obtained a head MAE mean ± std of 60 ± 22 HU and 47 ±11 HU, respectively. Therefore, exploring different loss functions is of interest because it can heavily affect intensity-linked errors.

Based on all the dosimetry results, very small discrepancies were obtained among the preprocessing applied. For instance, 3%/3 mm gamma indices equal to 99.83% ± 0.19% and 99.85% ± 0.17% were achieved for the experiments based on the combination of the HighResNet and WS standardization and optionally applying the N4 filter (Table 3). Although a significant P value of .012 was obtained, no major clinical impact is expected. As a result, this suggests that the proposed pCT generation method may

be suitable for introduction into clinics, regardless of the preprocessing applied.

The dose calculation algorithm used in this study was PB. An extra experiment was conducted to evaluate its relevance against Monte Carlo, considered more accurate in taking heterogeneities into account.[32,33] Bcause the latter is not commissioned in our institution for intensity modulated radiation therapy plans, we constituted an additional cohort of 8 brain tumor patients treated with arc therapy. Four out of 8 patients had a CT and a T1 MRI, the rest had a CT and a T1-Gd MRI. The preprocessing previously described in the Materials and Methods section was similarly applied, and the pCTs were generated. Dosimetry was performed on the pCT with the 2 different dose algorithms. No significant differences were observed for the DVH differences analysis ($P \geq .27$). A similar conclusion was obtained for the 3%/3 mm and 2%/2 mm gamma indices ($P \geq .40$). Concerning the 1%/1 mm criterion, 98.94% $\pm$ 0.68% and 98.40% $\pm$ 0.84% gamma pass rates were achieved for the PB and Monte Carlo algorithms, respectively ($P = .0078$). As a result, the PB approach is a reliable technique for the head localization owing to the absence of large inhomogeneities.

Regarding the data set, it was composed of 402 cases. To our knowledge, this is the largest cohort ever used in the head pCT generation field. Previous studies involved up to 77 patients.[12] Our data were split into independent sets: training, validation, and testing. Note that most of the published studies lack a validation set,[11,13,14,19,29,30] potentially leading to biased results.

MRI-only radiation therapy could remove isotropic 2 mm margins due to registration errors.[1] However, distortions can also lead to errors up to 2 mm even after applying a correction algorithm.[34] Therefore, establishing reliable quality assurance[35,36] is the key to unlock the full potential of radiation therapy.

Several limitations are present in this study. First, our DL pipeline necessitated paired images and thus intermodality registration, which can introduce errors in the training set. To evaluate this error, an experienced radiologist placed 3 landmarks both on the CT and the MRI of 10 patients. Registering the CT onto the MRI led to a mean distance error $\pm$ std of 3.0 mm $\pm$ 1.1 mm. Further investigation may focus on rigid registration errors and evaluate different algorithms, such as the FLIRT[37,38] tool, for comparison. Second, no analysis of the interplay effect of preprocessing steps and networks architecture was performed. Indeed, the use of a bias field correction and the selection of WS as the best standardization was based on experiments performed using HighResNet. This may have introduced bias in the comparison of HighResNet and 3D Unet.

Conclusions

In this study, we aimed at optimizing relevant parameters to achieve high-quality pCT for MR-only radiation therapy. The large variety of imaging devices and the considerable patient number constituting the training set appear to have a great impact on the pCT quality. All the parameters previously described, such as the MR sequence, intensity standardization, bias field correction, and network architecture, have a minor influence on dosimetry as the gamma indices and DVH differences remained clinically convincing for every technique in our cohort. This suggests the efficiency of the model and its possible introduction into clinics. Future work includes the extension of the current 3D network to integrate segmentation masks of target and organ-at-risk volumes and the development of a pCT generation model for a different anatomic site, such as the pelvis.

References

1. Ulin K, Urie MM, Cherlow JM. Results of a multi-institutional benchmark test for cranial CT/MR image registration. *Int J Radiat Oncol Biol Phys* 2010;77:1584-1589.
2. Schmidt MA, Payne GS. Radiotherapy planning using MRI. *Phys Med Biol* 2015;60:R323-R361.
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 Biol Phys* 2018;100:199-217.
4. Kang KM, Choi HS, Jeong BK, et al. MRI-based radiotherapy planning method using rigid image registration technique combined with outer body correction scheme: A feasibility study. *Oncotarget* 2017;8:54497-54505.
5. Wang C, Chao M, Lee L, et al. MRI-based treatment planning with electron density information mapped from CT images: A preliminary study. *Technol Cancer Res Treat* 2008;7:341-348.
6. 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.
7. Demol B, Boydev C, Korhonen J, et al. Dosimetric characterization of MRI-only treatment planning for brain tumors in atlas-based pseudo-CT images generated from standard T1-weighted MR images. *Med Phys* 2016;43:6557.
8. Fu J, Yang Y, Singhrao K, et al. Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic CT from MRI. *Med Phys* 2019;46:3788-3798.
9. Liu Y, Lei Y, Wang T, et al. MRI-based treatment planning for liver stereotactic body radiotherapy: Validation of a deep learning-based synthetic CT generation method. *Br J Radiol* 2019;92:20190067.
10. Liu Y, Lei Y, Wang Y, et al. MRI-based treatment planning for proton radiotherapy: Dosimetric validation of a deep learning-based liver synthetic CT generation method. *Phys Med Biol* 2019;64:145015.
11. Dinkla AM, Wolterink JM, Maspero M, et al. MR-only brain radiation therapy: Dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network. *Int J Radiat Oncol Biol Phys* 2018;102:801-812.
12. Kazemifar S, McGuire S, Timmerman R, et al. MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiother Oncol* 2019;136:56-63.
13. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys* 2017;44:1408-1419.
14. Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med Phys* 2018;45:3627-3636.
15. Reinhold JC, Dewey BE, Carass A, et al. Evaluating the impact of intensity normalization on MR image synthesis. *Proc SPIE Int Soc Opt Eng* 2019;10949:109493H.

16. Kläser K, Markiewicz P, Ranzini M, et al. Deep boosted regression for MR to CT synthesis. In: Gooya A, Goksel O, Oguz I, Burgos N, editors. Simulation and Synthesis in Medical Imaging. Cham, Switzerland: Springer International Publishing; 2018. p. 61-70.
17. Kläser K, Varsavsky T, Markiewicz P, et al. Improved MR to CT synthesis for PET/MR attenuation correction using imitation learning. In: Burgos N, Gooya A, Svoboda D, editors. Simulation and Synthesis in Medical Imaging. Cham, Switzerland: Springer International Publishing; 2019. p. 13-21.
18. Nie D, Trullo R, Lian J, et al. Medical image synthesis with context-aware generative adversarial networks. *Med Image Comput Comput Assist Interv* 2017;10435:417-425.
19. Wolterink JM, Dinkla AM, Savenije MHF, et al. Deep MR to CT synthesis using unpaired data. In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL, editors. Simulation and Synthesis in Medical Imaging. Cham, Switzerland: Springer International Publishing; 2017. p. 14-23.
20. Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. *Magn Reson Med* 1999;42:1072-1081.
21. Shinohara R, Sweeney E, Goldsmith J, et al. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin* 2014;6:9-19.
22. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp* 2019;40:4952-4964.
23. Li W, Wang G, Fidon L, et al. On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task. In: Niethammer M, Styner M, Aylward S, et al., editors. Information Processing in Medical Imaging. Cham, Switzerland: Springer International Publishing; 2017. p. 348-360.
24. Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, et al., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 424-432.
25. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization; 2016.
26. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29:1310-1320.
27. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979;9:62-66.
28. Mohan R, Chui C, Lidofsky L. Differential pencil beam dose computation model for photons. *Med Phys* 1986;13:64-73.
29. Liu F, Yadav P, Baschnagel AM, et al. MR-based treatment planning in radiation therapy using a deep learning approach. *J Appl Clin Med Phys* 2019;20:105-114.
30. Lei Y, Harms J, Wang T, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med Phys* 2019;46:3565-3581.
31. Chu C, Zhmoginov A, Sandler M. CycleGAN, a master of steganography. *ArXiv* 2017. abs/1712.02950.
32. Fragoso M, Wen N, Kumar S, et al. Dosimetric verification and clinical evaluation of a new commercially available Monte Carlo-based dose algorithm for application in stereotactic body radiation therapy (SBRT) treatment planning. *Phys Med Biol* 2010;55:4445-4464.
33. Petoukhova AL, van Wingerden K, Wiggenraad RGJ, et al. Verification measurements and clinical evaluation of the iPlan RT Monte Carlo dose algorithm for 6 MV photon energy. *Phys Med Biol* 2010; 55:4601-4614.
34. Weygand J, Fuller CD, Ibbott GS, et al. Spatial precision in magnetic resonance imaging-guided radiation therapy: The role of geometric distortion. *Int J Radiat Oncol Biol Phys* 2016;95:1304-1316.
35. Xing A, Holloway L, Arumugam S, et al. Commissioning and quality control of a dedicated wide bore 3T MRI simulator for radiotherapy planning. *Int J Cancer* 2016;4:421.
36. Sun J, Barnes M, Dowling J, et al. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. *Australas Phys Eng Sci Med* 2015;38:39-46.
37. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal* 2001;5:143-156.
38. Jenkinson M, Bannister P, Brady M, et al. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 2002;17:825-841.
39. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* 1998;17:87-97.
40. Cox IJ, Roy S, Hingorani SL. Dynamic histogram warping of image pairs for constant image brightness, *Proceedings of the 1995 International Conference on Image Processing (Vol2)-Volume 2 - Volume 2*. IEEE Computer Society; 1995;2366.