Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region

Hossein Arabi
Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva CH-1211, Switzerland

Jason A. Dowling
CSIRO Australian e-Health Research Centre, Herston, QLD, Australia

Ninon Burgos
Inria Paris, Aramis Project-Team, Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS UMR 7225 Sorbonne Université, Paris F-75013, France

Xiao Han
Elekta Inc., Maryland Heights, MO 63043, USA

Peter B. Greer
Calvary Mater Newcastle Hospital, Waratah, NSW, Australia
University of Newcastle, Callaghan, NSW, Australia

Nikolaos Koutsouvelis
Division of Radiation Oncology, Geneva University Hospital, Geneva CH-1211, Switzerland

Habib Zaidi[a)]
Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva CH-1211, Switzerland
Geneva University Neurocenter University of Geneva, Geneva 1205, Switzerland
Department of Nuclear Medicine and Molecular Imaging, University of Groningen, Groningen, the Netherlands
Department of Nuclear Medicine, University of Southern Denmark, Odense DK-500, Denmark

(Received 7 June 2018; revised 29 July 2018; accepted for publication 6 September 2018; published 10 October 2018)

Purpose: Magnetic resonance imaging (MRI)-guided radiation therapy (RT) treatment planning is limited by the fact that the electron density distribution required for dose calculation is not readily provided by MR imaging. We compare a selection of novel synthetic CT generation algorithms recently reported in the literature, including segmentation-based, atlas-based and machine learning techniques, using the same cohort of patients and quantitative evaluation metrics.

Methods: Six MRI-guided synthetic CT generation algorithms were evaluated: one segmentation technique into a single tissue class (water-only), four atlas-based techniques, namely, median value of atlas images (ALMedian)[1], atlas-based local weighted voting (ALWV)[2], bone enhanced atlas-based local weighted voting (ALWV-Bone)[3], iterative atlas-based local weighted voting (ALWV-Iter)[4], and a machine learning technique using deep convolution neural network (DCNN)[5].

Results: Organ auto-contouring from MR images was evaluated for bladder, rectum, bones, and body boundary. Overall, DCNN exhibited higher segmentation accuracy resulting in Dice indices (DSC) of 0.93 ± 0.17, 0.90 ± 0.04, and 0.93 ± 0.02 for bladder, rectum, and bones, respectively. On the other hand, ALMedian showed the lowest accuracy with DSC of 0.82 ± 0.20, 0.81 ± 0.08, and 0.88 ± 0.04, respectively. DCNN reached the best performance in terms of accurate derivation of synthetic CT values within each organ, with a mean absolute error within the body contour of 32.7 ± 7.9 HU, followed by the advanced atlas-based methods (ALWV: 40.5 ± 8.2 HU, ALWV-Iter: 42.4 ± 8.1 HU, ALWV-Bone: 44.0 ± 8.9 HU). ALMedian led to the highest error (52.1 ± 11.1 HU). Considering the dosimetric evaluation results, ALWV-Iter, ALWV, DCNN and ALWV-Bone led to similar mean dose estimation within each organ at risk and target volume with less than 1% dose discrepancy. However, the two-dimensional gamma analysis demonstrated higher pass rates for ALWV-Bone, DCNN, ALMedian and ALWV-Iter at 1%/1 mm criterion with 94.99 ± 5.15%, 94.59 ± 5.65%, 93.68 ± 5.53% and 93.10 ± 5.99% success, respectively, while ALWV and water-only resulted in 86.91 ± 13.50% and 80.77 ± 12.10%, respectively.

Conclusions: Overall, machine learning and advanced atlas-based methods exhibited promising performance by achieving reliable organ segmentation and synthetic CT generation. DCNN appears to have slightly better performance by achieving accurate automated organ segmentation and relatively small dosimetric errors (followed closely by advanced atlas-based methods, which in some cases achieved similar performance). However, the DCNN approach showed higher vulnerability to anatomical variation, where a greater number of outliers was observed with this method. Considering

the dosimetric results obtained from the evaluated methods, the challenge of electron density estimation from MR images can be resolved with a clinically tolerable error. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13187]

Key words: MRI-guided radiotherapy planning, CT synthesis, segmentation, atlas-based, machine learning

1. INTRODUCTION

Computed tomography (CT) is critical in radiotherapy treatment planning as there is a direct relationship between CT intensity values and tissue electron densities. However, organ delineation on CT images is challenging owing to low soft-tissue contrast, in particular for brain, head and neck, and pelvic regions[6]. In this regard, magnetic resonance imaging (MRI), as a structural imaging modality, offers improved soft-tissue contrast and organ visualization compared to CT images. Beside excellent soft-tissue contrast, advantages of MRI-guided radiation therapy (RT) planning include no exposure to ionizing radiation and cost reduction as no planning CT needs to be acquired, multi-parametric imaging possibilities offered by MRI, and elimination of uncertainties raised by imperfect coregistration.[7,8] The fact that MRI does not directly provide electron density information owing to different physical principles, challenges the calculation of dose distribution based on MRI and is, aside from engineering aspects, one of the major limitations of combining an MR scanner with a linear accelerator.[9]

Eliminating the CT scan from the radiation therapy planning chain is not a trivial problem and necessitates accurate estimation of electron density from the alternative MRI modality. The techniques proposed in the literature to estimate electron density maps from MR images can be categorized into three generic categories.[9–11] (a) *Bulk segmentation:* This approach relies on bulk segmentation of the MR images into a number of tissue classes followed by setting a homogenous predefined density to each region. Tissue classes commonly include water, fat, air, or bone.[12] Bony structures can be identified using a specialized ultra-short echo time (UTE) sequence, which is able to capture transient signal from components with a short T_2 relaxation time.[13–15] This sequence, however, suffers from long acquisition time compared to conventional MR sequences, low signal-to-noise ratio and partial volume effect which lead to significant bone segmentation errors.[16] (b) *Atlas-based methods:* This approach involves rigid and nonrigid mapping of atlas CT images onto a target MR image. An atlas dataset of aligned CT-MR image pairs is created to accurately map complex anatomy, where, in the first step, MR atlas images are mapped to the target MR image pairwise. Thereafter, the corresponding atlas CT images are transformed using these transformation maps and an estimate of the target electron density map is obtained by fusion of the transformed atlas CT images.[17,18] (c) *Machine learning:* Generally speaking, this approach attempts to estimate tissues electron densities directly from intensities of MR images. The training phase is the fundamental part of this approach which involves a data set of aligned CT-MR image pairs, similar to the atlas-based method. The algorithm learns the intensity mapping from MR image to electron density map normally through highly nonlinear systems. There are a number of highly effective machine learning algorithms such as random forest[19] and convolutional neural networks[5,20] which have shown excellent capability to estimate accurate electron density maps directly from MR images.

So far, many of the algorithms proposed in the literature and those implemented clinically have demonstrated promising performance and as such, they seem ready for use in clinical setting with acceptable errors.[21–23] Aside from the brain, MRI-guided radiation therapy for the prostate in the pelvic region has attracted much attention recently as some algorithms are becoming clinically available.[21,24,25] Nevertheless, there is a lack of direct comparison across the state-of-the-art methods and consensus on a scheme of evaluation reflecting the key performance parameters. The algorithms are commonly evaluated using different patient cohorts bearing different diseases and diverse image quality characteristics. A recent study (part of the *Gold Atlas project*) reported on the development of a dedicated MR-CT dataset with multi-observer delineations of organs designed for the assessment of synthetic CT generation in the pelvic region.[26] Comparison of different methods using the same cohorts, RT planning settings and evaluation standards gives valuable insight into robustness, potential applicability and expected quantitation errors in a clinical setting.

In this work, the aim is to assess the performance of multiple state-of-the-art synthetic CT generation methods for MRI-only radiation treatment planning in the pelvic region. A unified comparison between six MRI-based electron density estimation methods was conducted using a common cohort of MR-CT image pairs. Identical dosimetric metrics and reference CT-based RT plans were used to evaluate the accuracy and robustness of the methods. Invitations were sent out to major authors/groups who reported modern algorithms for synthetic CT generation from MRI. Special attention was paid to include techniques capable of automatic organ segmentation and avoid redundancy since some methods were quite similar. Authors/groups who accepted the invitation were involved in this comparative study. A related study compared a number of MRI-based synthetic CT generation methods in the pelvic region for prostate radiotherapy.[27] This study focused only on bulk density assignment (predefined electron density values) to the different tissue classes. The evaluated methods relied on manual bone (cortical and spongy) delineation, which renders their clinical relevance debatable. In our work, the selected methods exhibited

promising performance in addition to featuring fully automated CT synthesis and being capable of performing automated organ segmentation, which makes them suitable for clinical use.

At least one representative method from each of three generic synthetic CT generation categories has been included. As representative of the bulk segmentation approach, a two-class electron density map containing predefined values for water and air was generated from each MR image. Considering the number of publications, the atlas-based CT synthesis approach is highly popular and efficient for MRI-only RT planning.[1,4,28] For this generic method, four state-of-the-art atlas-based methods demonstrating promising results were chosen, namely, the median value of atlas images (ALMedian),[1] atlas-based local weighted voting (ALWV) proposed by,[2] bone enhanced atlas-based local weighted voting (ALWV-Bone),[3] and the iterative atlas-based local weighted voting (ALWV-Iter).[4] In the computer vision literature, deep learning and convolutional neural networks have demonstrated superior performance and are becoming the method of choice in many fields of computer vision as well as computerized medical image segmentation and identification problems.[29–31] Thus, we chose a machine learning method based on deep convolutional neural network in our evaluation. This neural network approach for synthetic CT generation (DCNN)[5] was originally developed for brain imaging to learn direct image to image conversion between MR and CT pairs. This method was adapted to pelvic images and evaluated along with other methods as a state-of-the-art machine learning method.

Aside from generating synthetic CT containing continuous or discretized attenuation properties of tissues, MRI-only RT planning can be further expanded if proper segmentation of the key organs at risk and target volumes can be performed automatically from MR images. The methods evaluated in this study are also able to perform automated organ segmentation from MR images. Thus, the organ delineation accuracy was assessed using standard segmentation metrics. Further validation was performed through comparison of doses calculated on the reference planning CT and synthetic CT images through a number of standard dosimetric analyses.

2. MATERIALS AND METHODS

2.A. Clinical data acquisition

All patient data sets were obtained retrospectively from Calvary Mater Newcastle Hospital, where the same data set was used previously to develop the automatic substitute CT generation (ALWV) method as described in Ref.[2] The cohort contains 39 patients aged between 58 and 78 years and body mass indices ranged from 19.1 to 35.4 kg/m^2. For each patient, three prostate pure gold fiducial markers of diameter 1.0 mm and length 3.0 mm were transrectally inserted before the planning image acquisition. CT scans (256 × 256 × 128 matrix with a voxel size of 1.5 mm × 1.5 mm × 2 (or 2.5 mm) were acquired on either a GE (Milwaukee, USA) LightSpeed RT or a Toshiba Aquilion (Tokyo, Japan) with slice thickness of 2.5 and 2.0 mm, respectively. Patients were scanned with a full bladder and empty rectum while positioned supine with knee and ankle immobilization stocks on a rigid couch-top.

The MR images were acquired with a Siemens (Erlangen, Germany) Skyra 3T scanner equipped for MR simulation with a dedicated radiation therapy flat couch. The planning MR sequence consists of a three-dimensional T2-weighted 1.6 mm isotropic SPACE (Sampling Perfection with Application optimized Contrast using different flip angle Evolution) with large field of view to cover the whole-pelvis area including the bladder. Organ contouring was performed by three experienced observers for bladder and rectum and the ground-truth contours were generated using majority voting to combine the observer decisions as described in a previous study.[32] The ground-truth contour for each organ was generated using majority voting to combine the observer decisions. Identification of bone from CT images was performed using the automatic segmentation tool implemented on Varian Eclipse and then, the obtained contour was rigidly transferred to whole-pelvis MR images and manually edited if required.

The cohort of 39 patients containing aligned whole-pelvis T2 and CT image pairs were used as input to train the different algorithms and to generate the synthetic CT images. Together with MR-CT image pairs, the delineated organs including the prostate, rectum, bones, and body contour, in the form of binary masks were available to examine automatic organ segmentation from MR images.

2.B. Synthetic CT generation algorithms

2.B.1. Median value of atlas images (ALMedian)

A simple multi-atlas approach was examined aiming to give an insight into how advanced methods performed differently compared to simpler methods. Given the series of CT-MR images aligned to the target MR images, the *ALMedian* synthetic CT is created by calculating the median value of the entire atlas CT images for each voxel independently using Eq. (1),[1] where x denotes the voxel index in the atlas CT number n (ACT_n):

$$ALMedain\ (x) = median(ACT_1(x),\ \ldots, ACT_n(x)) \qquad (1)$$

A major advantage of calculating the median value instead of the mean value for each voxel is that it shows greater robustness to outliers and the ability to better deal with multimodal distributions.

2.B.2. Atlas-based local weighted voting (ALWV)[2]

Atlas-based local weighted voting (ALWV) involves mapping information from a set of patient MRI and CT scans to a new patient MR image. The ALWV extends and improves on an earlier synthetic CT generation method[23], which used coupled group-wise (or average) atlases. The initial step in the ALWV method is the offline development

of an atlas set which contains image data from a large number of patients (ideally representing population anatomical variability). Each patient contributes an MR image; a CT scan which has been accurately coregistered to the MR image; and matching manual contours (prostate, bladder, rectum, bones and body).

The steps to convert a new MR image to synthetic CT are described in depth in Ref.[2] In brief the conversion process for a new MR image commences with an image preprocessing step to reduce artifacts and improve registration accuracy. Following this step each of the MR images in the atlas set is registered (rigid + nonrigid) to the target MR image in a pairwise manner. The rigid transform and nonrigid deformation field from each registration are then applied to the matching coregistered CT and contours from the atlas set.

Once the registrations are completed a small 3D patch around each voxel in the set of registered MR images is compared with a patch from the same spatial location on the new MR image. Local weighted voting[33] is used on each patch to provide a measure of MR intensity similarity between the registered images and the new MR image (more similar patches receive higher weights). After normalization, this weighting is applied to the same voxels in the coregistered CT scans and these values are combined to generate HU value estimates for each synthetic CT voxel. The same weights are also applied to the coregistered contours, generating automatic contours which are useful for quality assurance purposes. CT synthesis is finalized by a 1-mm expansion to the synthetic CT to account for the missing outer skin layer.

2.B.3. Bone enhanced atlas-based local weighted voting (ALWV-Bone)[3]

The original algorithm was developed based on in-phase Dixon MR images and for this study the same framework was adapted to T2-weighted MR image. Initially, the T2-weighted MR images were corrected for magnetic field inhomogeneity, noise and inter-image MR intensity nonuniformity. After preprocessing, MR images in the atlas dataset were registered to the target MR image through a leave-one-out-cross-validation (LOOCV) scheme with a combination of rigid and nonrigid registration based on normalized mutual information and B-spline interpolator using Elastix open source software as described previously.[34] Then, the atlas CT images were mapped to the target MR image using the obtained transformation maps.

In the first step, bone segmentation was performed on the target MR image through voxel-by-voxel atlas voting scheme. This step leads to a binary bone map which can be assumed to represent the most likely bone delineation of the target image and helps to achieve atlas fusion with special emphasis on the bony structures. Considering MR_n and BL_n denote aligned atlas MR images and corresponding bone label maps, respectively, the bone segmentation (Bt) of the target image (Tr) can be performed using:

$$\widehat{Bt}(x) = arg\max_L \sum_{n=1}^{N} p_n((Tr(x)|MR_n(x))) \; p_n((Bt(x)|BL_n(x))) \quad (2)$$

where N is the number of training subjects in the atlas dataset. Estimation of the target bone at each voxel ($\widehat{Bt}$(x)) depends on image morphology likelihood $p_n(Tr(x), MR_n(x))$ between the target and the atlas MR images as well as bone label prior $p_n(Bt(x), BL_n(x))$. Phase congruency map (PCM) was used to calculate image morphology likelihood, which is robust to inter-subject intensity variation and noise while providing valuable structural information.[35] The bone label prior ($p_n(Bt(x), BL_n(x))$) was calculated based on the signed distance transform from the bone label maps (BL_n). The output of this step ($\widehat{Bt}$) is used in the next step to define weighting factors for each atlas image through comparison of the signed distanced maps calculated on $\widehat{Bt}$ and each of atlas bone maps (BL_n). The continuous valued synthetic CT images were generated using a voxel-wise weighted atlas fusion framework based on the PCM morphology likelihood as well as the calculated signed distance maps.

2.B.4. Iterative atlas-based local weighted voting (ALWV-Iter)[4]

The iterative multi-atlas propagation framework developed by[4] combines in a single pipeline segmentation and CT synthesis. The method relies on a multi-atlas database consisting for each atlas of coregistered structural MR image A^{MRI}, CT image A^{CT}, and a set of manually segmented images A^S. The dataset of the n^{th} atlas is denoted by $\mathbb{A}_n = \{A_n^{MRI}, A_n^{CT}, A_n^S\}$.

At the initial iteration ($t = 1$), the target subject's dataset is only composed of the subject's MR image: $\mathrm{L}_1 = \{I^{MRI}\}$. A set of probabilistic segmentations and synthetic CT image is jointly generated from this target MR image by registering each atlas MR image A_n^{MRI} to the target MR image I^{MRI}, and fusing the propagated atlas segmentations and CT images according to the similarity between each atlas MR image A_n^{MRI} and the target MR image I^{MRI}.

For the subsequent iterations, the previously generated set of probabilistic segmentations I_{t-1}^S and pCT image I_{t-1}^{pCT} is combined with the target MR image I^{MRI} to generate a new target subject's dataset $\mathrm{L}_t = \{I^{MRI}, I_{t-1}^{pCT}, I_{t-1}^S\}$. A new refined set of probabilistic segmentations and pCT image is jointly generated from the new target dataset L_t, first by registering each atlas dataset $\mathbb{A}_n$ to the target dataset L_t (multichannel registration). The propagated atlas segmentations and CT images are then fused, not only according to the similarity between each atlas MR image A_n^{MRI} and the target MR image I^{MRI}, but according to the similarity between each atlas dataset $\mathbb{A}_n$ and the target dataset L_t. As a compromise between accuracy and computation complexity, the process is stopped after the fourth iteration.

2.B.5. Deep Convolutional Neural Network (DCNN)[5]

The synthetic CT for each subject was generated using a deep convolutional neural network method as described in

Ref.[5] The particular deep convolutional neural network model architecture is a slightly modified version of the *U-Net* model[36] that was originally proposed for image segmentation, where the encoding path is modified to match the first 13 layers of the popular 16-layer Visual Geometry Group (VGG) model.[37] This modification allows the encoding path parameters to be initialized using the pretrained VGG model to implement transfer learning. The decoding path is a mirrored version of the encoding path, with an extra 1×1 convolutional layer added in the end to map each 64-component feature vector from the previous layer to a CT number. In total, there are 27 convolutional layers in the model and 35 million model parameters.

To generate the synthetic CT for every subject, a fourfold cross-validation procedure is applied. The 39 subjects are divided into four groups. At each time, one group is retained as the test set, and the remaining three groups are used as training data to train the DCNN model. The training of the method was originally performed using 18 brain patient scans and a sixfold cross-validation scheme.[5] The larger sample size of the training datasets used in this work is sufficient to train the DCNN. In addition, simple data augmentation was also adopted to artificially increase the number of datasets during model training. Once the model is trained, it is applied on each test subject's MR image to generate the corresponding synthetic CT. It is noted that the DCNN model works in 2D in that it takes a 2D MR axial slice as input and outputs the corresponding 2D slice of the final synthetic CT. Training the DCNN model takes about 2.5 days of computation time using a single NVIDIA Titan X GPU card. Once the model is trained, it takes approximated 9 s to process all axial slices of a new MR image to get the final 3D synthetic CT result. No pre- or post- processing is performed except that the N3 bias field correction algorithm[38] is applied on each MR image to reduce intensity nonuniformity artifacts. Bias field correction was also applied in ALMedian, ALWV and ALWV-Bone methods.

It should be noted that organ segmentations performed by the DCNN algorithm is a separate process from synthetic CT generation since the training of the DCNN and auto-segmentation are repeated for each organ separately. Conversely, organ segmentation in atlas-based methods is linked to the synthetic CT generation process. In these techniques, synthetic CT generation entails a voxel-wise weighting strategy to define the contribution weight of each voxel of atlas CT images based on the local similarity between the target MRI and the corresponding atlas MR images. The final synthetic CT is then generated through a voxel-wise weighted averaging of the atlas images. Likewise, organ segmentation is performed using voxel-wise weighting factors obtained from CT synthesis to weight the binary image of the organ delineation in the atlas dataset. Then, for each voxel, the binary vote of each atlas image is weighted and averaged (to create a probability organ map with a threshold of 0.5) to define the organ volume on the target MRI.

2.B.6. Water-only

The dose distributions were also calculated for the water-only synthetic CT. To generate this electron density map, the body contour was segmented from the target MR image followed by assigning Hounsfield Unit of water (HU = 0) to all voxels within the body contour and HU = −1000 to the background air.

2.C. Evaluation strategy

2.C.1. Image segmentation accuracy metrics

Automatic organ delineations from MR images generated by different methods were compared against the ground-truth contours defined manually. The organs included in the automatic segmentation evaluation were bladder, rectum, bone, and body contour. In addition to automatic bone delineation, bone segmentation directly from generated synthetic CTs using the intensity threshold of 140 HU (referred to Bone-thresh in result section) was evaluated separately. The segmentation accuracy was evaluated using the Dice similarity coefficient (DSC) [Eq. (3)], which represents the overlap of the two volume divided by the total volume of the two objects, and the mean absolute surface distance (MASD), which measures the average of the absolute Euclidean distance (d_{ave}) between two segmentation surfaces ($S_{R,A}$) [Eq. (4)].

$$DSC(A,R) = \frac{2|A \cap R|}{|A| + |R|} \tag{3}$$

$$MASD(A,R) = \frac{d_{ave}\ (S_A,\ S_R)\ +\ d_{ave}\ (S_R,\ S_A)}{2} \tag{4}$$

where A represents the automated segmentation volume and R denotes the reference organ delineation.

2.C.2. Synthetic CT generation accuracy

In addition to volumetric evaluation of automatic organ segmentation, the mean error (ME) and mean absolute error (MAE) were calculated between reference CTs (R_{CT}) and synthetic CTs (A_{CT}) for each organ (namely, bladder, rectum, bones, Bone-thresh, and inside body contour) taking into account all of the voxels within the segmentation volume (P) as follows:

$$MAE_{CT} = \frac{1}{P}\sum_{i=1}^{P}|A_{CT}(i) - R_{CT}(i)| \tag{5}$$

$$ME_{CT} = \frac{1}{P}\sum_{i=1}^{P}A_{CT}(i) - R_{CT}(i) \tag{6}$$

Moreover, linear regression analysis and joint histograms of MRI-derived synthetic CTs versus reference CT images were performed for each technique averaged over 39 subjects.

The entire voxel within the body contour including the margins and the edges of the contours were included in ME and MAE calculation as well as linear regression analysis.

2.C.3. Dosimetric evaluation

The Eclipse™ treatment planning system (Varian Medical Systems Inc, Palo Alto, CA, USA) was employed for treatment planning based on the volumetric-modulated arc therapy (VMAT) technique. Conventional RT treatment planning was performed to deliver 36.25 Gy dose to the planning target volume (PTV). The PTV provided a safety margin of 10 mm around the clinical target volumes, which included the prostate and seminal vesicles. 3D dose distributions were calculated for the PTV and organs at risk (OARs), which included the bladder, rectum, left, and right heads of femur (HOF). It should be noted that automated organs segmentation was not used for RT planning as OARs and target volumes delineations were performed manually.

The radiation therapy plans optimized on the reference CTs (reference dose matrix) were copied onto the synthetic CT images and dose distributions were recalculated accordingly. Dose volume histograms (DVH), representing a histogram in a 2D graphical format relating radiation dose to organ volume, were exported for the target volumes and OARs. The dose calculation was carried out using the anisotropic analytical algorithm (AAA v. 13) using a 6 MV photon beam and a dose matrix of $2.5 \times 2.5 \times 2.5$ mm^3.

The comparison between synthetic CT and reference CT dose distribution maps was performed by calculating minimum (Dose-min), maximum (Dose-max) and mean (Dose-mean) absorbed doses for different OARs and target volumes. Then, Eqs. 7 and 8 were used to calculate the mean and mean absolute absorbed dose errors for each synthetic CT generation algorithm across the 39 subjects. Here, $Dose_{CT}$ and $Dose_{pCT}$ denote planned doses calculated on the reference CT and synthetic CT images, respectively.

$$ME_{Dose}\ (\%) = 100 \times \frac{Dose_{pCT} - Dose_{CT}}{Dose_{CT}} \tag{7}$$

$$MAE_{Dose}\ (\%) = 100 \times \frac{\left|Dose_{pCT} - Dose_{CT}\right|}{Dose_{CT}} \tag{8}$$

The dose distributions recalculated over the different synthetic CT images were also evaluated against the reference CT dose in terms of the absolute volume (cc) receiving a certain level of dose discrepancy. To this end, a voxel-by-voxel dose difference map was computed and used to measure the associated volume having a dose discrepancy equal or greater than a certain dose difference (for instance at a dose level of 1 Gy, the total volume bearing a dose difference equal to or greater than 1 Gy is reported). The dose discrepancy levels were plotted versus the corresponding accumulated volumes ranging from −1 to 1 Gy.

Furthermore, a two-dimensional Gamma analysis,[39] which is a commonly used metric for comparing the dose distributions combining features of dose difference and distance-to-agreement, was employed to analyze the axial dose distributions intersecting the treatment isocenter. The Gamma indices were presented for 3%/3 mm, 2%/2 mm and 1%/1 mm dose difference/distance-to-agreement. The collimator angle was 30° and all the dose levels were included taking the local (pixel) value of the CT dose map as reference. The skin dose was not taken into account as well as doses below 10% of the prescribed dose.

Paired t-test analysis with a significance level of 0.05 was performed to assess if the differences between the obtained results are statistically significant.

3. RESULTS

Representative slices of the different synthetic CT images are presented in Fig. 1 together with the target MR and reference CT images. The binary bone maps obtained from applying intensity threshold of 140 HU on electron density maps are shown next to each CT image. Visual inspection reveals comparable bone extraction achieved by the different synthetic CT generation methods.

As mentioned earlier, the methods evaluated in this work are capable of performing automated organ contouring from MR images. The results of the quantitative evaluation of the automated organ contouring are summarized in Table I for bladder, rectum, bone, body, and bone extracted from the CT images. The DSC index and MASD are averaged over the 39 patients. The mean difference and absolute mean difference between synthetic CT and reference CT images within each automatically obtained contours from MR images are presented as well. In general, DCNN exhibited better automated organ delineation and more accurate CT value estimation followed by ALWV-Iter and ALWV-Bone, and ALWV. The joint histogram analysis operating over the 39 patients demonstrated a similar trend in Fig. 2. DCNN achieved the closest estimation of the CT values followed by ALWV-Bone and ALWV-Iter. DCNN achieved the lowest root mean square error (RMSE), comparing all voxels values in synthetic CT and reference CT images, of 2.33 HU while 3.23, 3.39, 4.59, 3.55, and 11.48 HU achieved by ALWV-Iter, ALWV-Bone, ALWV, ALMedian, and water-only, respectively.

The therapy doses recalculated over the synthetic CT images were measured within each body organ and target volume, then the average errors (relative mean ± SD) of Dose_min, Dose_max and Dose_mean were computed using Eqs. (7) and (8). Tables II and III show the results corresponding to the different organs at risk and target volumes. The mean absorbed doses within the organs at risk and target volumes did not exhibit large difference between ALWV-Bone, ALWV-Iter, ALWV, and DCNN methods, however, DCNN achieved slightly lower errors. The statistical analysis did not reveal statistically significant differences between the mean absorbed doses within the different organs when using the various synthetic CT generation methods (Tables II and III). Given the dose maps, representative slices of dose distributions calculated for the different methods together with

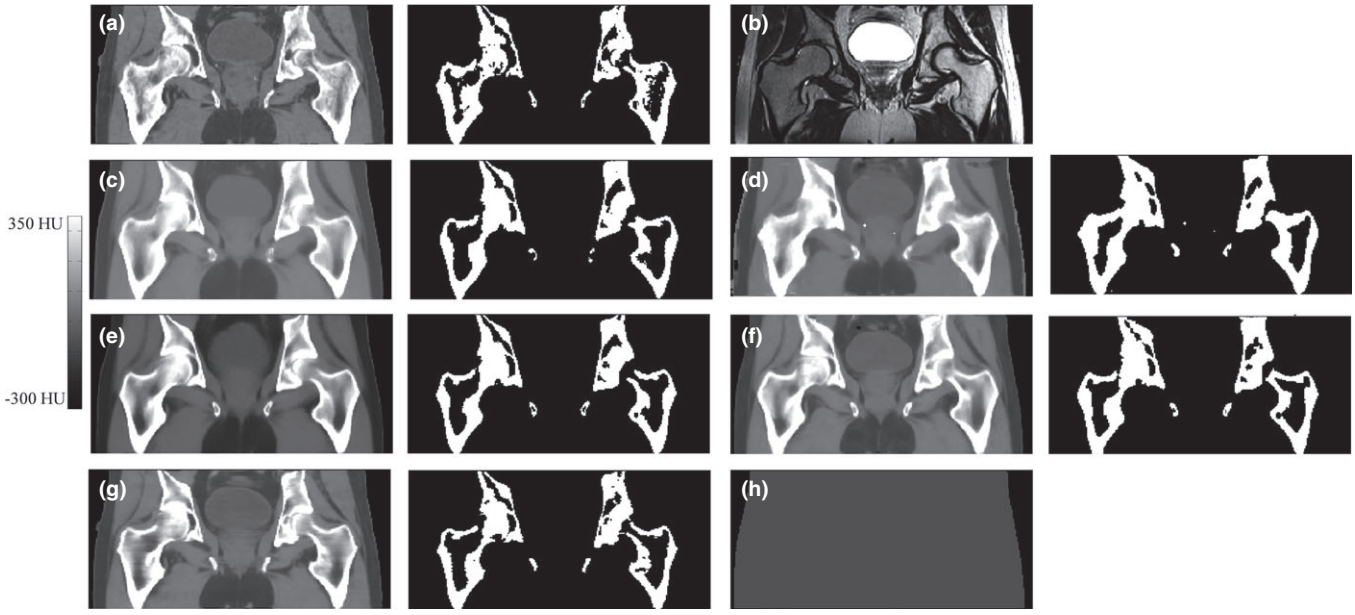

FIG. 1. Representative slice of MRI-derived synthetic CT images together with a binary map of bone tissue segmented through applying an intensity threshold of 140 HU. (a) Reference CT, (b) Reference MR image, (c) Synthetic CT of ALMedian, (d) ALWV, (e) ALWV-Bone, (f) ALWV-Iter, (g) DCNN, and h) water-only.

TABLE I. Accuracy of automatic organ contouring from MR images using different synthetic CT generation methods compared with the reference CT images and manual organ delineation on MR images in terms of Dice similarity (DSC), mean absolute surface distance (MASD), mean error (ME) and mean absolute error (MAE) of Hounsfield unit (HU) calculated within each automated contour (mean ± SD). Bone-thresh: bone segmented by applying intensity threshold of 140 HU. *P*-values are also shown.

Bladder	ALMedian	ALWV	ALWV-Bone	ALWV-Iter	DCNN
DSC	0.82 ± 0.20 ($P < 0.01$)	0.86 ± 0.12 ($P < 0.01$)	0.86 ± 0.13 ($P < 0.01$)	0.87 ± 0.10 ($P < 0.01$)	0.93 ± 0.17 ($P < 0.01$)
MASD (mm)	7.01 ± 4.17 ($P < 0.01$)	5.10 ± 4.57 ($P = 0.01$)	7.88 ± 4.78 ($P < 0.01$)	7.56 ± 4.42 ($P < 0.01$)	2.36 ± 2.44 ($P = 0.02$)
ME (HU)	−1.5 ± 20.2 ($P = 0.51$)	−2.9 ± 18.7 ($P = 0.43$)	8.1 ± 17.6 ($P = 0.12$)	7.7 ± 16.1 ($P = 0.1$)	−1.8 ± 12.9 ($P = 0.43$)
MAE (HU)	30.0 ± 17.6 ($P < 0.01$)	24.1 ± 13.6 ($P < 0.01$)	26.4 ± 12.7 ($P < 0.01$)	25.2 ± 10.1 ($P < 0.01$)	18.4 ± 6.6 ($P < 0.01$)
Rectum	**ALMedian**	**ALWV**	**ALWV-Bone**	**ALWV-Iter**	**DCNN**
DSC	0.81 ± 0.08 ($P < 0.01$)	0.84 ± 0.06 ($P < 0.01$)	0.84 ± 0.70 ($P < 0.01$)	0.84 ± 0.06 ($P < 0.01$)	0.90 ± 0.04 ($P < 0.01$)
MASD (mm)	5.03 ± 2.72	2.37 ± 1.34	4.95 ± 2.39	4.81 ± 2.22	2.09 ± 1.11
ME (HU)	37.6 ± 84.9 ($P = 0.02$)	6.9 ± 81.7 ($P = 0.18$)	27.6 ± 90.5 ($P = 0.03$)	−30.3 ± 94.6 ($P = 0.03$)	22.7 ± 84.8 ($P = 0.05$)
MAE (HU)	93.5 ± 71.2 ($P < 0.01$)	88.1 ± 60.8 ($P < 0.01$)	100.0 ± 62.0 ($P < 0.01$)	114.8 ± 63.6 ($P < 0.01$)	78.3 ± 69.2 ($P = 0.01$)
Body	**ALMedian**	**ALWV**	**ALWV-Bone**	**ALWV-Iter**	**DCNN**
DSC	0.98 ± 0.02 ($P = 0.01$)	1.00 ± 0.00 ($P < 0.01$)	0.99 ± 0.01 ($P = 0.02$)	0.99 ± 0.01 (P = 0.02)	0.99 ± 0.01 ($P = 0.03$)
MASD (mm)	4.12 ± 01.89 ($P < 0.01$)	0.55 ± 0.56 ($P = 0.04$)	3.01 ± 1.86 ($P = 0.03$)	3.93 ± 2.98 ($P = 0.03$)	1.78 ± 0.63 ($P = 0.04$)
ME (HU)	10.2 ± 18.3 ($P = 0.31$)	−0.6 ± 14.2 ($P = 0.61$)	8.7 ± 15.6 ($P = 0.27$)	2.0 ± 15.1 ($P = 0.49$)	3.5 ± 11.7 ($P = 0.43$)
MAE (HU)	52.1 ± 11.1 ($P < 0.01$)	40.5 ± 8.2 ($P < 0.01$)	44.0 ± 8.9 ($P < 0.01$)	42.4 ± 8.1 ($P < 0.01$)	32.7 ± 7.9 ($P < 0.01$)
Bone	**ALMedian**	**ALWV**	**ALWV-Bone**	**ALWV-Iter**	**DCNN**
DSC	0.88 ± 0.04 ($P < 0.01$)	0.91 ± 0.03 ($P < 0.01$)	0.92 ± 0.02 ($P < 0.01$)	0.92 ± 0.02 ($P < 0.01$)	0.93 ± 0.02 ($P < 0.01$)
MASD (mm)	3.73 ± 0.58 ($P < 0.01$)	1.45 ± 0.47 ($P = 0.01$)	1.94 ± 0.45 ($P = 0.01$)	2.07 ± 0.43 ($P < 0.01$)	3.51 ± 3.92 ($P < 0.01$)
ME (HU)	−32.9 ± 55.4 ($P = 0.03$)	−6.4 ± 46.5 ($P = 0.36$)	26.6 ± 56.7 ($P = 0.04$)	19.5 ± 46.3 ($P = 0.09$)	−4.1 ± 40.7 ($P = 0.43$)
MAE (HU)	161.1 ± 30.0 ($P < 0.01$)	134.2 ± 24.0 ($P < 0.01$)	163.8 ± 25.0 ($P < 0.01$)	130.2 ± 23.4 ($P < 0.01$)	119.9 ± 22.6 ($P < 0.01$)
Bone-thresh	**ALMedian**	**ALWV**	**ALWV-Bone**	**ALWV-Iter**	**DCNN**
DSC	0.79 ± 0.06 ($P < 0.01$)	0.81 ± 0.06 ($P < 0.01$)	0.84 ± 0.05 ($P < 0.01$)	0.83 ± 0.05 ($P < 0.01$)	0.83 ± 0.05 ($P < 0.01$)
MASD (mm)	3.67 ± 1.71 ($P < 0.01$)	2.63 ± 0.51 ($P = 0.01$)	2.94 ± 1.48 ($P = 0.01$)	2.16 ± 0.23 ($P = 0.01$)	2.12 ± 0.22 ($P = 0.01$)
ME (HU)	−35.4 ± 59.7 ($P < 0.01$)	16.0 ± 51.0 ($P = 0.09$)	40.3 ± 60.0 ($P < 0.01$)	26.2 ± 54.2 ($P = 0.02$)	14.4 ± 46.9 ($P = 0.08$)
MAE (HU)	151.4 ± 31.3 ($P < 0.01$)	143.6 ± 24.8 ($P < 0.01$)	172.9 ± 25.8 ($P < 0.01$)	138.8 ± 24.9 ($P < 0.01$)	127.3 ± 26.3 ($P < 0.01$)

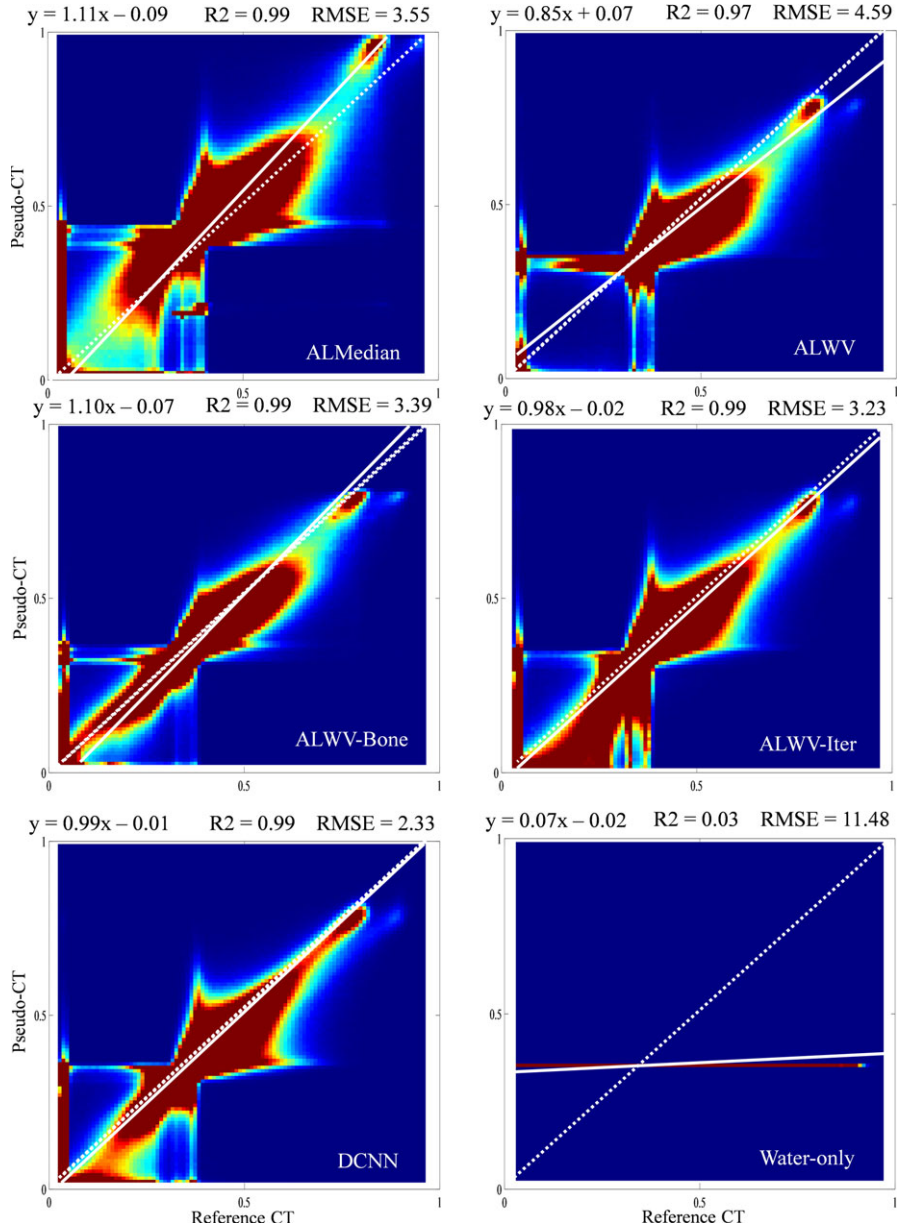

FIG. 2. Joint histograms averaged across 39 subjects, between the reference CT and synthetic CT images generated using different methods. Images are min/max scaled between 0 and 1 [Color figure can be viewed at wileyonlinelibrary.com]

dose distribution error map are provided in Fig. 3. Moreover, similar to Fig. 2, the joint histogram analysis was performed over the dose distribution maps, comparing the plans recalculated on synthetic CT images to reference CT, are illustrated in Fig. 4. ALWV-Iter achieved the lowest RMSE (1.15) compared to other techniques, which resulted in a RMSE of 1.60, 1.73, 1.74, 2.47, and 4.78 for ALMedian, ALWV-Bone, DCNN, ALWV, and water-only, respectively.

A representative graph of DVH obtained from recalculating the absorbed doses over different synthetic CTs are presented in Figs. 5 and 6 for some OARs (bladder, right femur, left femur, and rectum) and target volumes (CTV, PTV, and PTV_3 mm), respectively. Point-by-point analyses over the DVH graphs are presented in the supplemental material.

Table IV summarizes two-dimensional gamma analysis of isocenter dose distributions for three dose difference/distance-to-agreement criteria (3%/3 mm, 2%/2 mm and 1%/1 mm). Considering the 3%/3 mm criterion, all methods performed similarly with more than 98% (P <0.01) pass rate. Likewise, at a pass threshold of 2%/2 mm, deep learning and atlas-based algorithms performed equally well (>96.93%). Water-only lagged behind with a success rate of 95.38% (P <0.01). Lowering the pass threshold to 1%/1 mm results in an apparent difference between the different methods where ALWV-Bone, DCNN, ALMedian, and ALWV-Iter achieved a pass rate of more than 93% followed by ALWV and Water-only with 86.91% (P <0.01) and 80.77% (P <0.01), respectively. In addition to gamma analysis, the accumulative

TABLE II. Average error (relative mean(%) ± SD) of Dose_min, Dose_max and Dose_mean for plans recalculated on different synthetic CT images calculated for the following organs at risks.

Bladder	ALMedian ME ± SD MAE ± SD	ALWV ME ± SD MAE ± SD	ALWV-Bone ME ± SD MAE ± SD	ALWV-Iter ME ± SD MAE ± SD	DCNN ME ± SD MAE ± SD	Water-only ME ± SD MAE ± SD
Dose_min	0.74 ± 5.49	1.32 ± 5.40	0.98 ± 5.45	0.97 ± 5.40	1.01 ± 5.49	0.13 ± 5.12
	1.36 ± 5.36	1.52 ± 5.35	1.35 ± 5.37	1.35 ± 5.31	1.39 ± 5.41	1.79 ± 4.79
Dose_max	0.15 ± 0.75	−0.76 ± 0.77	−0.1 ± 0.72	0.15 ± 0.78	−0.03 ± 0.69	0.54 ± 1.18
	0.63 ± 0.42	0.85 ± 0.67	0.57 ± 0.43	0.64 ± 0.45	0.51 ± 0.46	0.96 ± 0.87
Dose_mean	0.09 ± 0.66	−0.55 ± 0.76	−0.05 ± 0.6	0.14 ± 0.68	−0.02 ± 0.57	−0.06 ± 1.10
	0.52 ± 0.41	0.75 ± 0.56	0.47 ± 0.37	0.53 ± 0.44	0.41 ± 0.39	0.82 ± 0.73

Rectum	ALMedian ME ± SD MAE ± SD	ALWV ME ± SD MAE ± SD	ALWV-Bone ME ± SD MAE ± SD	ALWV-Iter ME ± SD MAE ± SD	DCNN ME ± SD MAE ± SD	Water-only ME ± SD MAE ± SD
Dose_min	0.05 ± 1.12	−0.08 ± 1.06	0.18 ± 0.99	−0.39 ± 1.22	−0.13 ± 1.00	−0.09 ± 1.82
	0.80 ± 0.77	0.76 ± 0.73	0.70 ± 0.70	0.88 ± 0.93	0.74 ± 0.68	1.36 ± 1.20
Dose_max	0.20 ± 0.80	−0.76 ± 0.77	−0.03 ± 0.75	0.18 ± 0.84	−0.05 ± 0.74	0.74 ± 1.38
	0.67 ± 0.46	0.85 ± 0.66	0.59 ± 0.45	0.70 ± 0.49	0.52 ± 0.52	1.22 ± 0.97
Dose_mean	0.28 ± 0.78	-0.59 ± 0.74	0.09 ± 0.72	0.15 ± 0.82	0.06 ± 0.73	0.47 ± 1.45
	0.66 ± 0.49	0.70 ± 0.63	0.56 ± 0.46	0.63 ± 0.53	0.50 ± 0.52	0.98 ± 1.16

Left HOF	ALMedian ME ± SD MAE ± SD	ALWV ME ± SD MAE ± SD	ALWV-Bone ME ± SD MAE ± SD	ALWV-Iter ME ± SD MAE ± SD	DCNN ME ± SD MAE ± SD	Water-only ME ± SD MAE ± SD
Dose_min	0.47 ± 1.02	0.10 ± 0.98	−0.09 ± 1.13	−0.22 ± 1.19	0.1 ± 0.94	4.89 ± 2.13
	0.82 ± 0.77	0.76 ± 0.61	0.78 ± 0.81	0.88 ± 0.82	0.68 ± 0.65	4.89 ± 2.13
Dose_max	0.03 ± 0.72	−0.76 ± 0.78	−0.18 ± 0.73	0.10 ± 0.76	−0.1 ± 0.63	1.62 ± 1.10
	0.60 ± 0.39	0.89 ± 0.62	0.61 ± 0.42	0.60 ± 0.47	0.49 ± 0.40	1.62 ± 1.10
Dose_mean	0.06 ± 0.49	−0.67 ± 0.56	−0.14 ± 0.48	0.07 ± 0.48	−0.07 ± 0.43	1.27 ± 0.69
	0.40 ± 0.27	0.72 ± 0.50	0.40 ± 0.30	0.38 ± 0.31	0.32 ± 0.30	1.27 ± 0.69

Right HOF	ALMedian ME ± SD MAE ± SD	ALWV ME ± SD MAE ± SD	ALWV-Bone ME ± SD MAE ± SD	ALWV-Iter ME ± SD MAE ± SD	DCNN ME ± SD MAE ± SD	Water-only ME ± SD MAE ± SD
Dose_min	0.52 ± 0.99	0.17 ± 0.93	0.14 ± 0.89	−0.04 ± 0.88	0.18 ± 0.86	4.72 ± 2.01
	0.83 ± 0.74	0.68 ± 0.65	0.68 ± 0.58	0.71 ± 0.50	0.64 ± 0.60	4.72 ± 2.01
Dose_max	0.10 ± 0.68	−0.61 ± 0.74	−0.14 ± 0.68	0.11 ± 0.69	−0.07 ± 0.60	1.72 ± 1.02
	0.56 ± 0.40	0.76 ± 0.58	0.53 ± 0.44	0.53 ± 0.45	0.45 ± 0.40	1.72 ± 1.02
Dose_mean	0.09 ± 0.48	−0.63 ± 0.58	−0.14 ± 0.48	0.04 ± 0.49	−0.08 ± 0.42	1.29 ± 0.67
	0.40 ± 0.28	0.69 ± 0.51	0.39 ± 0.31	0.39 ± 0.30	0.32 ± 0.28	1.29 ± 0.67

volume associated with a certain dose difference threshold is presented in Fig. 7. The volume corresponding to each dose difference level is the average value measured over the entire number of patients.

In addition to the evaluation of methods using conventional global and local metrics, a number of cases where the methods failed to produce proper segmentation or generation of synthetic CT images are documented in supplemental Figs. 1–4. Three cases presenting with bladder segmentation failure when using DCNN (two cases) and ALWV-Iter (one case) are reported in supplemental Fig. S1. ALMedian and ALWV failed to identify bony structures for one patient (supplemental Fig. S2). Incomplete electron density map and body contour were observed for ALWV-Bone, DCNN, and ALWV in one patient (supplemental Figs. S3 and S4). These cases were included in the analysis of the results.

4. DISCUSSION

The primary aim of this work was to assess the performance of a number of state-of-the-art MRI-only radiation planning methods using multiple dosimetric and segmentation metrics evaluated across a common patient cohort. The intention was to evaluate not only a representative approach of each generic type but also focusing on highly promising approaches proposed in the literature to provide valuable

TABLE III. Average error (relative mean(%) ± SD Dev) of Dose_min, Dose_max and Dose_mean for plans recalculated on different synthetic CT images calculated for the following targets.

CTV	ALMedian ME ± SD MAE ± SD	ALWV ME ± SD MAE ± SD	ALWV-Bone ME ± SD MAE ± SD	ALWV-Iter ME ± SD MAE ± SD	DCNN ME ± SD MAE ± SD	Water-only ME ± SD MAE ± SD
Dose_min	0.25 ± 0.76	−0.68 ± 0.78	0.00 ± 0.75	0.20 ± 0.72	0.01 ± 0.72	0.85 ± 1.30
	0.67 ± 0.44	0.85 ± 0.59	0.61 ± 0.43	0.58 ± 0.46	0.53 ± 0.48	1.21 ± 0.96
Dose_max	−2.72 ± 0.97	−2.18 ± 1.15	−2.98 ± 0.93	−2.67 ± 0.85	−2.91 ± 0.88	−4.84 ± 1.85
	2.72 ± 0.97	2.23 ± 1.06	2.98 ± 0.93	2.67 ± 0.85	2.91 ± 0.88	4.84 ± 1.85
Dose_mean	0.22 ± 0.73	−0.73 ± 0.73	−0.05 ± 0.69	0.24 ± 0.74	−0.02 ± 0.67	0.86 ± 1.21
	0.62 ± 0.42	0.83 ± 0.61	0.55 ± 0.42	0.63 ± 0.43	0.51 ± 0.43	1.14 ± 0.95
PTV	**ALMedian ME ± SD MAE ± SD**	**ALWV ME ± SD MAE ± SD**	**ALWV-Bone ME ± SD MAE ± SD**	**ALWV-Iter ME ± SD MAE ± SD**	**DCNN ME ± SD MAE ± SD**	**Water-only ME ± SD MAE ± SD**
Dose_min	0.33 ± 0.95	−0.62 ± 0.93	0.09 ± 0.94	0.14 ± 1.13	0.07 ± 1.07	0.84 ± 1.62
	0.75 ± 0.66	0.88 ± 0.68	0.68 ± 0.64	0.86 ± 0.74	0.71 ± 0.79	1.25 ± 1.32
Dose_max	−0.58 ± 1.29	−1.03 ± 1.33	−0.85 ± 1.26	−0.56 ± 1.30	−0.76 ± 1.36	−0.65 ± 2.18
	1.04 ± 0.94	1.27 ± 1.09	1.11 ± 1.03	1.02 ± 0.96	1.11 ± 1.10	1.69 ± 1.50
Dose_mean	0.23 ± 0.72	−0.72 ± 0.71	−0.04 ± 0.68	0.18 ± 0.72	−0.01 ± 0.64	0.82 ± 1.22
	0.61 ± 0.44	0.81 ± 0.60	0.54 ± 0.41	0.60 ± 0.42	0.47 ± 0.43	1.10 ± 0.97
PTV_3 mm	**ALMedian ME ± SD MAE ± SD**	**ALWV ME ± SD MAE ± SD**	**ALWV-Bone ME ± SD MAE ± SD**	**ALWV-Iter ME ± SD MAE ± SD**	**DCNN ME ± SD MAE ± SD**	**Water-only ME ± SD MAE ± SD**
Dose_min	0.27 ± 0.96	−0.67 ± 0.93	0.05 ± 0.92	0.18 ± 1.02	0.05 ± 0.96	0.66 ± 1.60
	0.74 ± 0.65	0.89 ± 0.71	0.66 ± 0.64	0.75 ± 0.70	0.62 ± 0.73	1.17 ± 1.27
Dose_max	−2.48 ± 1.12	−2.10 ± 1.17	−2.65 ± 1.39	−2.46 ± 1.02	−2.67 ± 1.10	−4.43 ± 2.10
	2.51 ± 1.06	2.15 ± 1.09	2.80 ± 1.05	2.48 ± 0.99	2.71 ± 1.00	4.54 ± 1.85
Dose_mean	0.21 ± 0.73	−0.73 ± 0.72	−0.04 ± 0.69	0.18 ± 0.73	−0.03 ± 0.65	0.78 ± 1.26
	0.62 ± 0.44	0.83 ± 0.61	0.55 ± 0.41	0.60 ± 0.44	0.49 ± 0.43	1.11 ± 0.97

insight into accuracy and range of bias expected in MRI-only radiation therapy. In this regard, four newly proposed methods, namely, ALWV, ALWV-Bone, ALWV-Iter, and DCNN, which exhibited excellent results in the literature, were evaluated in this work.

Automated organ segmentation from MR images can further facilitate the introduction of MRI-only RT into the clinic. Among the atlas-based methods, ALWV-Iter and ALWV-Bone combined organ segmentation and synthetic CT estimation in a single pipeline with the aim to improve both the accuracy of segmentation and synthetic CT estimation. The machine learning algorithm performs organ segmentation in a more customized manner as for the DCNN method, in particular, the network training was repeated for each organ individually which justifies the overall better segmentation performance of this approach. DCNN exhibited superior performance particularly for the bladder and rectum auto-segmentation with a DSC of 0.93 and 0.90, respectively, compared to ALWV-Iter as the second-best approach, which achieved a DSC of 0.87 and 0.84, respectively. However, ALWV-Bone and ALWV achieved almost similar results as ALWV-Iter (DSC of 0.86 and 0.84, respectively). In the ALWV-Iter approach, the probabilistic segmentation and synthetic CT generation are optimized simultaneously, as opposed to DCNN where the training is repeated for each organ individually, to benefit the synergy of the joint estimation. ALWV-Bone method relies on the similar idea to ALWV-Iter's but only concentrating on bone tissue, which justifies its slightly better performance in bone identification (Bone-thresh).

Aside from volumetric evaluation of auto-contouring, the assessment of CT values estimation within each organ revealed slightly more accurate performance of the DCNN approach as it produced the smallest MAEs for all organs reported in Table I. For instance considering the body contour, DCNN resulted in MAE of 32.7 ± 7.9 HU followed by ALWV, ALWV-Iter, ALWV-Bone and ALMedian with MAEs of 40.5 ± 8.2, 42.4 ± 8.1, 44.0 ± 8.9, and 52.1 ± 11.1, respectively (P <0.03). The ALWV method led to accurate synthetic CT value estimation within organs, however, the overall dose calculation is not as good. This can be partially justified by the addition of a skin margin to the final synthetic CT images, which was carried out in this method to account for missing MR signal. The machine learning algorithms in general and deep convolutional neural networks in particular are capable of grasping a highly nonlinear MR intensity to CT value correspondence,[5] which explains better performance of DCNN compared to the other atlas-based in

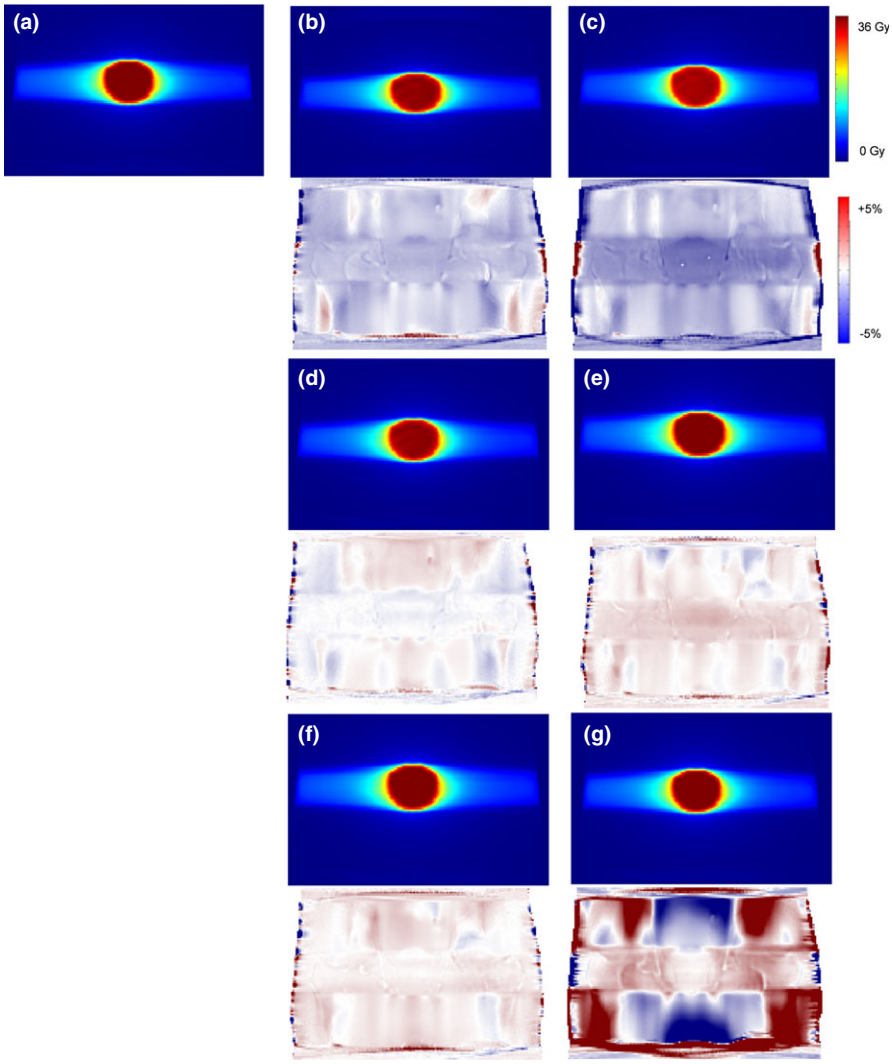

FIG. 3. Representative slices of dose distribution calculated for different synthetic CT maps together with dose distribution error map. (a) Reference CT, (b) ALMedian, (c) ALWV, (d) ALWV-Bone, (e) ALWV-Iter, (f) DCNN, and (g) Water-only. [Color figure can be viewed at wileyonlinelibrary.com]

this regard. On the other hand, the atlas-based techniques evaluated in this work rely mostly on the similarity of structures between target and atlas MR images to define the fusion weights, which may not be as accurate as DCNN to estimate correct synthetic CT values.

Considering the mean absorbed dose differences measured for the different OARs and target volumes in Tables II and III, aside from water-only approach, all other methods resulted in less than 1% mean absorbed dose error as well as mean absolute error. The largest mean absorbed dose difference for water-only occurred in bony tissue (right HOF), which can be simply justified by the absence of bone in the electron density map. However, largest errors of mean absorbed dose for other methods were observed in the target volumes. However, the differences between the mean absorbed doses within all organs were not statistically significant (P >0.3). The results presented in Tables II and III are consistent with earlier evaluation of ALWV-Bone, ALWV, and ALWV-Iter approaches performed on different datasets and study settings.[2–4]

Gamma analysis, which quantifies the point-by-point difference between measured and calculated dose distributions in terms of both distance-to-agreement and dose discrepancies, has become the gold standard metric for the comparison between measured and calculated absorbed dose distributions. A pass rate of 1%/1 mm demonstrated comparable performance of ALWV-Bone, DCNN, ALMedian, and ALWV-Iter. Moreover, the difference between the different techniques became evident when the pass rate was lowered from 2%/2 mm to 1%/1 mm where ALWV drops to ~87% with a large standard deviation while the others stay in the 90s%. Although a pass rate greater than 93% at 1%/1 mm was achieved using atlas-based and DCNN methods, relatively large standard deviations (>5) implies that at smaller scales these methods may fail to reach this accuracy. Although, the gamma index is a standard metric for dose verification in RT planning quality assurance, the interactions between the parameters used in the gamma index calculation (the distance-to-agreement and dose difference) complicates the interpretation of the

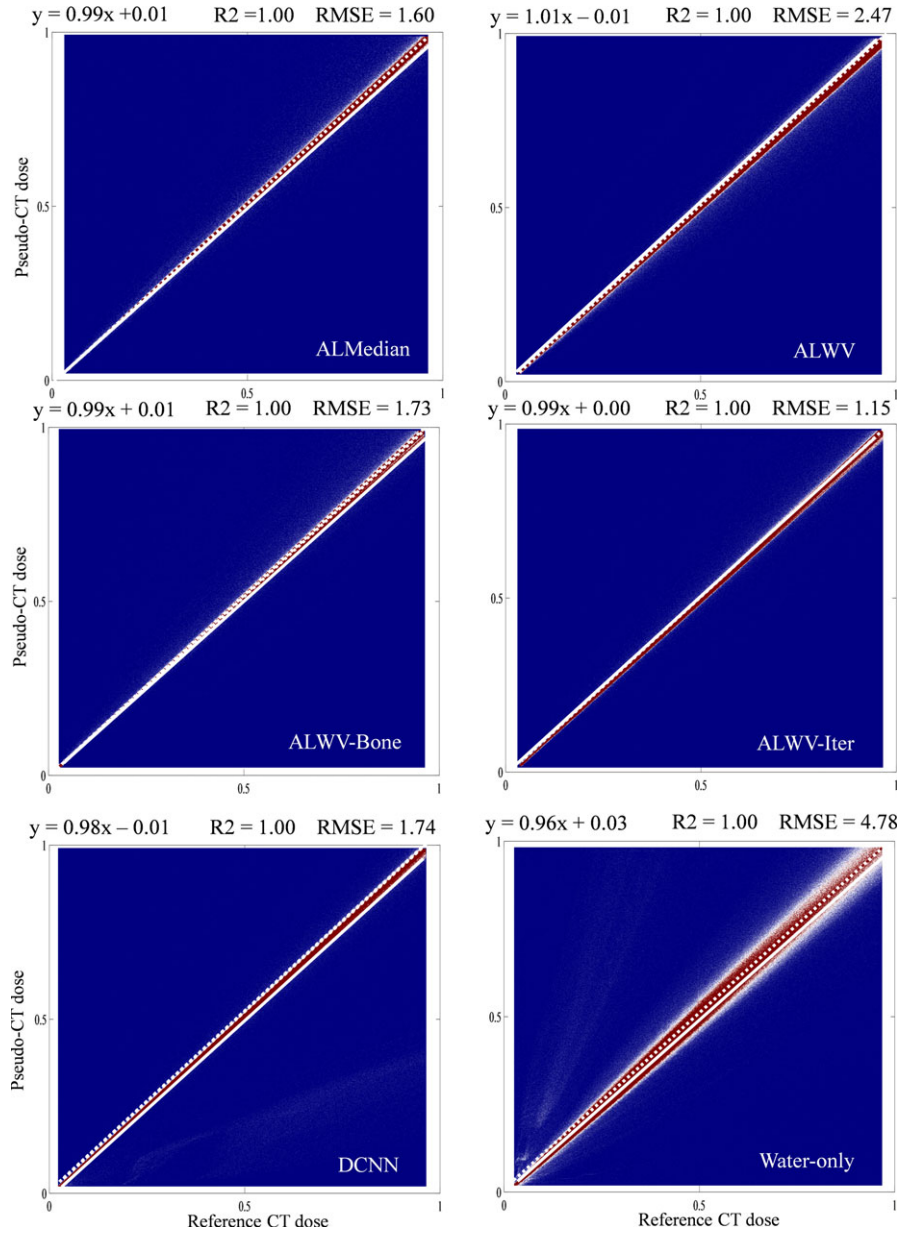

FIG. 4. Joint histograms averaged across 39 subjects, between the reference dose maps obtained from reference CT and synthetic CT images generated using different methods. Images are min/max scaled between 0 and 1. [Color figure can be viewed at wileyonlinelibrary.com]

outcome. In this regard, dose discrepancy versus volume analysis was performed to provide complementary information about the extent of absorbed dose difference (Fig. 7). Given the prescribed dose of 36.25 Gy, the volume corresponding to ±1 Gy dose difference (which indicates the total volume bearing a dose difference ≥±1 Gy) is almost equivalent to 3% of prescribed dose where no significant margin is observed between the different methods. On the other hand, considering a dose difference of ±0.4 Gy (which corresponds to almost 1% of the prescribed dose), larger margins are observed between the different methods but still showing comparable dose conformity, except for the water-only approach. In addition, a slightly positive bias is observed for ALMedian, ALWV, and ALWV-Bone techniques.

In addition to assessment of methods performed both globally and locally using multiple segmentation and dosimetric metrics, their robustness was examined through exploration of the outliers and observed gross defects. Supplemental Figs. S1–S4 illustrate the prominent cases related to both organ segmentation and synthetic CT generation. Considering the segmentation and dosimetric results reported so far, DCNN exhibited excellent overall performance in organ segmentation as well as electron density estimation for RT planning. Nevertheless, it displayed greater sensitivity to the outliers where in some cases, it performed poorly in organ segmentation and electron density map generation. Due to the fact that atlas-based methods, ALMedian, ALWV, ALWV-bone, and ALWV-Iter, rely on prior knowledge provided by the templates during the course of segmentation or electron density

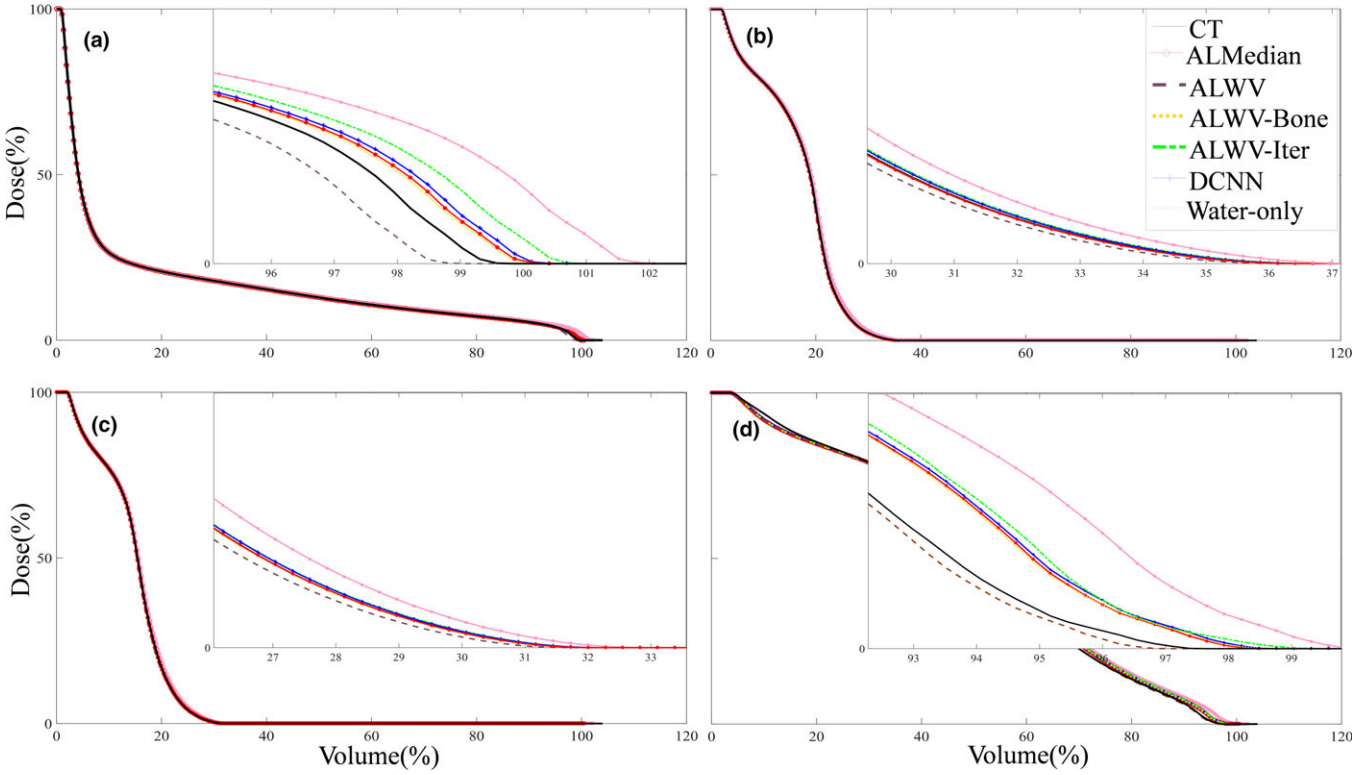

FIG. 5. Representative dose volume histogram of plans recalculated on different synthetic CT maps for different organs at risk. (a) Bladder, (b) Right femur, (c) Left femur, and (d) Rectum. [Color figure can be viewed at wileyonlinelibrary.com]

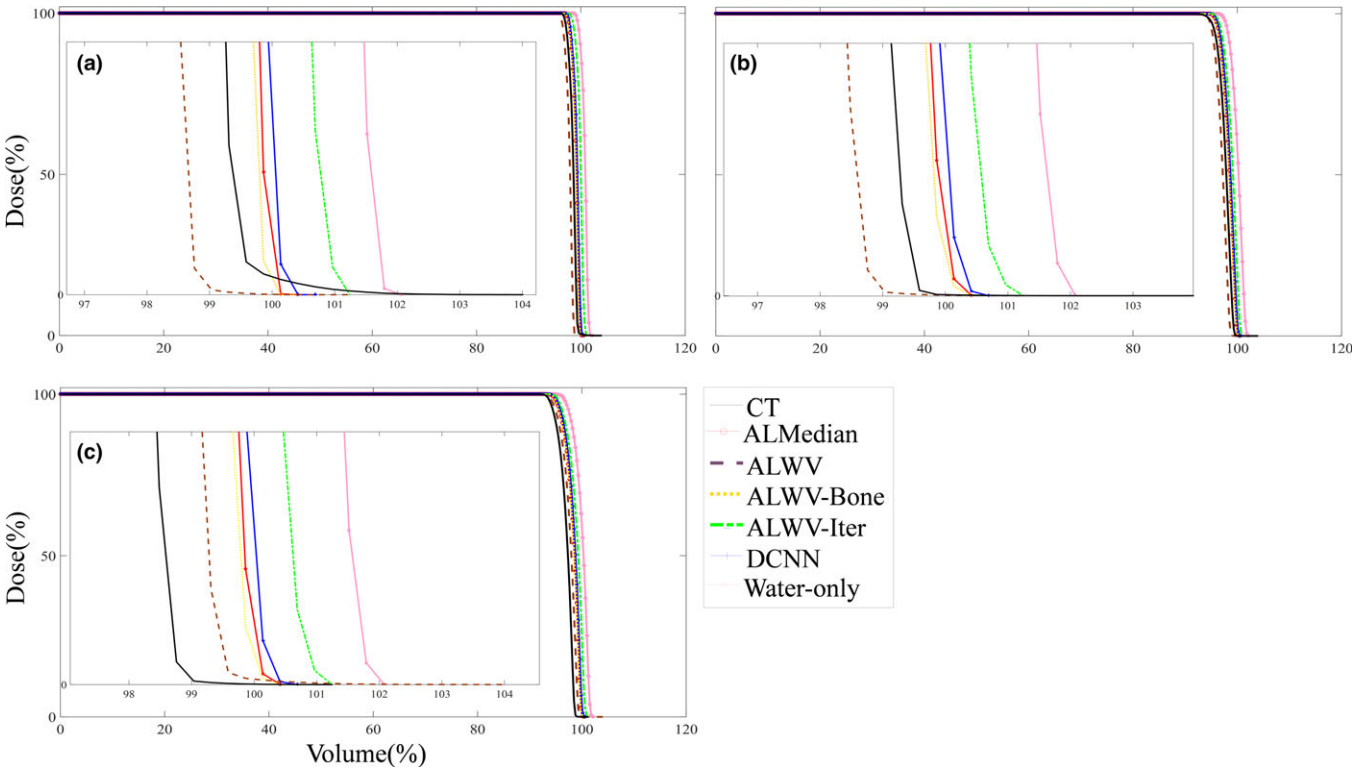

FIG. 6. Representative dose volume histogram of plans recalculated on different synthetic CT maps for target regions. (a) CTV, (b) PTV), and (c) PTV_3 mm. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE IV. Gamma analysis results comparing the original CT plan with the same plan recalculated on the synthetic CT images. The differences are statistically significant with a $P < 0.01$.

	ALMedian (mean ± SD)	ALWV (mean ± SD)	ALWV-Bone (mean ± SD)	ALWV-Iter (mean ± SD)	DCNN (mean ± SD)	Water-only (mean ± SD)
3%/3 mm	98.96 ± 0.78	98.41 ± 1.56	99.51 ± 0.32	98.96 ± 0.57	99.22 ± 0.46	98.22 ± 1.75
2%/2 mm	97.92 ± 1.49	96.93 ± 2.69	98.84 ± 0.48	97.99 ± 1.02	98.47 ± 0.68	95.38 ± 5.17
1%/1 mm	93.68 ± 5.53	86.91 ± 13.50	94.99 ± 5.15	93.10 ± 5.99	94.59 ± 5.65	80.77 ± 12.10

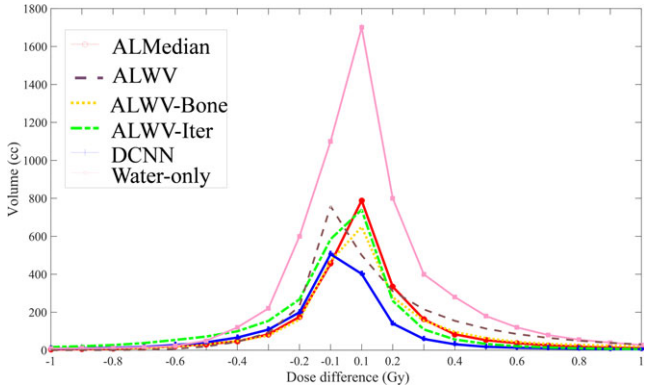

FIG. 7. Volumes (cc) associated with a certain magnitude of dose difference (Gy) between dose distributions calculated using reference CT and different synthetic CT images. [Color figure can be viewed at wileyonlinelibrary.com]

generation, they showed greater robustness to odd cases even when MR images suffered from artifacts or poor signal-to-noise ratio. In particular, advanced atlas-based methods which are able to select or give weight to more similar templates, are less likely to fail to perform proper segmentation. The use of a larger dataset containing a wide range of anatomical variations during the training of machine learning approaches would enhance the robustness of these methods. The major advantage of atlas-based methods is their robustness to variability in MR image quality and presence of artifacts as they rely on the prior knowledge of atlas images. In the case of gross artifacts or low MR image quality, atlas-based methods are capable of at least generating a synthetic CT image representing the average atlas population, which in most cases should not lead to a large quantitation error. Increasing the size of the atlas database can add to the robustness of atlas-based methods, however, the computation time rises linearly, thus impacting the clinical feasibility of the approach. Atlas registration is commonly performed in a pairwise manner, and as such, the generation of a synthetic CT might take a couple of hours. On the other hand, machine learning algorithms are less robust to variability in image quality and the presence of artifacts, but are capable of swift generation of a synthetic CT (couple of minutes), which facilitates their implementation in the clinic. Moreover, machine learning algorithms can be retrained over time by adding new samples to the training dataset to boost their robustness without affecting computational cost.

Overall, atlas-based methods exhibit relatively high robustness to image quality. However, this is highly dependent on the registration algorithm and atlas selection strategy employed. Most atlas-based methods, including those evaluated in this work, use robust registration and atlas selection algorithms, which makes them less sensitive to image quality. Besides, atlas-based methods exhibited higher robustness to outliers as they usually result in a realistic outcome representing the average of the atlas dataset. Advanced atlas-based methods (in particular ALWV-Bone and ALWV-Iter) showed even higher resilience to outliers owing to sophisticated atlas selection procedures. Moreover, they produced competitive segmentation performance (compared to DCNN) since CT synthesis and organ segmentations are jointly carried out to take into account the mutual dependency existing between these two variables. Although the DCNN method achieved the highest accuracy in organ segmentation and CT synthesis, it was less robust to the outliers in comparison with advanced atlas-based techniques. However, considering the computation time, DCNN is a more appealing choice for use in the clinic since the synthetic CT generation process takes less 1 min as opposed to atlas-based methods which take up to 2 hours. This is an important issue in MRI-guided adaptive RT planning (real-time adaptive replanning) involving re-optimization and creation of new treatment plans during RT fractions.[40]

One of the limitations of this study is that all subjects were acquired on the same 3T MRI scanner with the same image acquisition protocol. As such, the sensitivity of the algorithms to scanner types, setup practices, such as immobilization devices and field strengths were not assessed. In addition, the datasets did not contain any metal imposed artifact to particularly evaluate the methods robustness in the presence of severe image artifacts; however, this issue warrants further investigation. The introduction of the fast acquisition MR sequences such as ultra-short (UTE) and zero echo-times (ZTE), enabled automatic separation of bone and air[41–43] and therefore can be considered as an alternative image modality in RT planning. Evaluation of fast MR sequences in MRI-only RT planning compared to the advanced atlas-based and machine learning approaches would be very much appreciated particularly in the presence of metallic implants where UTE sequences are potentially more robust against metal artifacts.

5. CONCLUSION

State-of-the-art MRI-based synthetic CT generation methods were evaluated in the context of MRI-only RT planning using multiple segmentation, CT synthesis and dosimetric

metrics. We aimed to provide a comparative assessment of existing promising methods and demonstrate clinical feasibility in the pelvic region. The algorithm relying on deep convolutional neural network approach (DCNN) exhibited promising organ segmentation accuracy for bladder, bone, and rectum. However, atlas-based techniques (ALWV, ALWV-Bone, ALWV-Iter) showed comparable performance. Considering the mean absorbed dose in different OAR and target volumes, aside from water-only, all the other methods performed similarly, achieving less than 1% mean absorbed dose error. However, gamma analysis revealed that ALWV-Bone, DCNN, ALMedian, and ALWV-Iter demonstrated higher performance particularly at the 1%/1 mm criteria achieving an average passing rate of more than 93% with relatively large standard deviations (>5), which requires caution at small scales. Finally, the DCNN approach exhibited higher vulnerability to anatomical variation as it resulted in a larger number of outliers. Based on these results, it can be concluded that the challenges of MRI-only radiation therapy in the pelvic region are solvable with a clinically tolerable error.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation under grant No. SNSF 320030_176052 and the Swiss Cancer Research Foundation under Grant KFS-3855-02-2016. N.B. receives funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: habib.zaidi@hcuge.ch; Telephone: +41 22 372 7258; Fax: +41 22 372 7169.

REFERENCES

1. Sjölund J, Forsberg D, Andersson M, Knutsson H. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Phys Med Biol*. 2015;60:825.
2. 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.
3. Arabi H, Koutsouvelis N, Rouzaud M, Miralbell R, Zaidi H. 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.
4. Burgos N, Guerreiro F, McClelland J, et al. Iterative framework for the joint segmentation and CT synthesis of MR images: application to MRI-only radiotherapy treatment planning. *Phys Med Biol*. 2017;62:4237–4253.
5. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys*. 2017;44:1408–1419.
6. Rasch C, Steenbakkers R, van Herk M. Target definition in prostate, head, and neck. *Semin Radiat Oncol*. 2005;15:136–145.
7. Nyholm T, Jonsson J. Counterpoint: opportunities and challenges of a magnetic resonance imaging-only radiotherapy work flow. *Semin Radiat Oncol*. 2014;24:175–180.
8. Owrangi AM, Greer PB, Glide-Hurst CK. MRI-only treatment planning: benefits and challenges. *Phys Med Biol*. 2018;63:05TR01.
9. 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.
10. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol*. 2017;12:28.
11. Mehranian A, Arabi H, Zaidi H. Vision 20/20: magnetic resonance imaging-guided attenuation correction in PET/MRI: challenges, solutions, and opportunities. *Med Phys*. 2016;43:1130–1155.
12. Chin AL, Lin A, Anamalayil S, Teo BKK. Feasibility and limitations of bulk density assignment in MRI for head and neck IMRT treatment planning. *J Appl Clin Med Phys*. 2014;15:100–111.
13. Keereman V, Fierens Y, Broux T, De Deene Y, Lonneux M, Vandenberghe S. MRI-based attenuation correction for PET/MRI using ultrashort echo time sequences. *J Nucl Med*. 2010;51:812–818.
14. Edmund JM, Kjer HM, Van Leemput K, Hansen RH, Andersen JA, Andreasen D. A voxel-based investigation for MRI-only radiotherapy of the brain using ultra short echo times. *Phys Med Biol*. 2014;59:7501.
15. Wiesinger F, Bylund M, Yang J, et al. Zero TE-based pseudo-CT image conversion in the head and its application in PET/MR attenuation correction and MR-guided radiation therapy planning. *Magn Reson Med*. 2018;80:1440–1451.
16. Rank CM, Tremmel C, Hunemohr N, Nagel AM, Jakel O, Greilich S. MRI-based treatment plan simulation and adaptation for ion radiotherapy using a classification-based approach. *Radiat Oncol*. 2013;8:51.
17. Arabi H, Zaidi H. One registration multi-atlas-based pseudo-CT generation for attenuation correction in PET/MRI. *Eur J Nucl Med Mol Imaging*. 2016;43:2021–2035.
18. Burgos N, Cardoso M, Thielemans K, et al. Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies. *IEEE Trans Med Imaging*. 2014;33:2332–2341.
19. Huynh T, Gao Y, Kang J, et al. Estimating CT image from MRI data using structured random forest and auto-context model. *IEEE Trans Med Imaging*. 2016;35:174–183.
20. 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.
21. Tyagi N, Fontenla S, Zhang J, et al. Dosimetric and workflow evaluation of first commercial synthetic CT software for clinical use in pelvis. *Phys Med Biol*. 2017;62:2961–2975.
22. 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.
23. 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.
24. Persson E, Gustafsson C, Nordstrom 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 Biol Phys*. 2017;99:692–700.
25. Siversson C, Nordstrom F, Nilsson T, et al. Technical Note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm. *Med Phys*. 2015;42:6090–6097.
26. Nyholm T, Svensson S, Andersson S, et al. MR and CT data with multi observer delineations of organs in the pelvic area–part of the Gold Atlas project. *Med Phys*. 2018;45:1295–1300.
27. Kim J, Garbarino K, Schultz L, et al. Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy. *Radiat Oncol*. 2015;10:239.
28. Uh J, Merchant TE, Li Y, Li X, Hua C. MRI-based treatment planning with pseudo CT generated through atlas registration. *Med Phys*. 2014;41:051711–051718.

29. Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA, Išgum I. *Deep MR to CT Synthesis Using Unpaired Data*. Imaging: International Workshop on Simulation and Synthesis in Medical; 2017:14–23.
30. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
31. Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst C. Generating Synthetic CT s from Magnetic Resonance Images using Generative Adversarial Networks. *Med Phys*. 2018;45:3627–3636.
32. Chandra SS, Dowling JA, Shen K-K, et al. Patient specific prostate segmentation in 3-D magnetic resonance images. *IEEE Trans Med Imaging*. 2012;31:1955–1964.
33. Artaechevarria X, Munoz-Barrutia A, Ortiz-de-Solorzano C. Combination strategies in multi-atlas image segmentation: application to brain MR data. *IEEE Trans Med Imaging*. 2009;28:1266–1277.
34. Akbarzadeh A, Gutierrez D, Baskin A, et al. Evaluation of whole-body MR to CT deformable image registration. *J Appl Clin Med Phys*. 2013;14:238–253.
35. Kovesi P. Phase congruency: a low-level image invariant. *Psychol Res*. 2000;64:136–148.
36. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention Conference – MICCAI 2015, Cham, 2015; 234–241.
37. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ARXIV, arXiv preprint arXiv:1409. 1556. 2014.
38. 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.
39. Jonsson JH, Johansson A, Soderstrom K, Asklund T, Nyholm T. Treatment planning of intracranial targets on MRI derived substitute CT data. *Radiother Oncol*. 2013;108:118–122.
40. Pathmanathan AU, van As NJ, Kerkmeijer LGW, et al. Magnetic resonance imaging-guided adaptive radiation therapy: a "game changer" for prostate treatment? *Int J Radiat Oncol Biol Phys*. 2018;100:361–373.
41. Jonsson JH, Akhtari MM, Karlsson MG, Johansson A, Asklund T, Nyholm T. Accuracy of inverse treatment planning on substitute CT images derived from MR data for brain lesions. *Radiother Oncol*. 2015;10:13.
42. Yang Y, Cao M, Kaprealian T, et al. Accuracy of UTE-MRI-based patient setup for brain cancer radiation therapy. *Med Phys*. 2016;43:262–267.
43. Hsu S-H, Cao Y, Lawrence TS, et al. Quantitative characterizations of ultrashort echo (UTE) images for supporting air–bone separation in the head. *Phys Med Biol*. 2015;60:2869.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Three outliers in bladder segmentation. DCNN and ALWV-Iter automated contouring failed to identify bladder boundary in three cases, DCNN in cases 1 and 2 and ALWV-Iter in case 3. (A) Ground-truth bladder segmentation. (B) Automated bladder contouring. (C) Target MR image.

Fig. S2. Two outliers in bone identification. ALWV and ALMedian resulted in false bone identification in case 1 and 2, respectively. (A) MRI-derived pseudo-CT image together with bone map (obtained from intensity thresholding of 140 HU). B) Reference CT and bone map.

Fig. S3. Two outliers in pseudo-CT generation. ALWV-Bone and DCNN resulted in flawed electron density map in case 1 and 2, respectively. (A) Reference CT. (B) MRI-derived pseudo-CT. (C) Difference HU error map. (D) Reference dose distribution. (E) Recalculated dose distribution using the generated pseudo-CT image. (F) Difference dose error map.

Fig. S4. An outlier in body contour delineation and pseudo-CT generation. ALWV resulted in flawed electron density map for Case 2 shown in Supplemental Figure 3. (A) Reference CT (transaxial, sagittal and coronal views). (B) MRI-derived pseudo-CT. (C) Difference HU error map. (D) Reference dose distribution. E) Recalculated dose distribution using the generated pseudo-CT image. (F) Difference dose error map.

Table S1. Dosimetric errors [relative mean ± Std Dev (absolute mean ± Std Dev)] for the organs at risk calculated using total number of DVH points between D100% and D0% in dose increments of 0.1 Gy.

Table S2. Dosimetric errors [relative mean ± Std Dev (absolute mean ± Std Dev)] for target regions calculated using total number of DVH points between D100% and D0% in dose increments of 0.1 Gy.